Title Page

Title of Project: GestureClean: A Touchless Interaction Language for the Operating Room

Principal Investigator:

- Juan P. Wachs, Professor, School of Industrial Engineering, Purdue University.

Co-Investigators:

- Richard Rodgers, MD, Assistant Professor of Clinical Neurosurgery at Indiana University School of Medicine.
- Linsong Zhang, PhD, Assistant Professor, Department of Statistics, Purdue University.
- Lisa Goffman, PhD, Professor of Speech, Language, and Hearing Sciences, and Professor of Linguistics.

Team Members:

1. Naveen Madapana, PhD student; School of Industrial Engineering, Purdue University.

Organization: Purdue University

Inclusive Dates of Project: 09/01/2016 – 08/31/2019

Federal Project Officer: David Rodrick

Acknowledgement of Agency Support: This project was supported by grant number 1R18HS024887 from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the AHRQ. The AHRQ had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Grant Award Number: 1R18HS024887

Structured Abstract

Purpose: Asepsis requirements are critical in the Operating Room (OR) and need to be fulfilled to avoid the spread of nosocomial infections and any risk of contamination. This work systematically developed a touchless interaction system for the OR to allow surgeons to control systems in a sterile manner.

Scope: Touchless interfaces powered by gestures and speech allow surgeons to control medical imaging systems autonomously while maintaining total asepsis in the OR. The choice of best gestures/commands for such interfaces is a critical step that determines the overall efficiency of surgeon-computer interaction. In this regard, this work proposes three different ways of obtaining best gestures: 1. gesture-elicitation; 2. qualitative; and 3. usability approaches. A conjugation of these three approaches was used to design and implement our touchless system in the OR.

Methods: We hypothesize that there is a correlation between gestures' qualitative properties (v) and their usability metrics (u). We conducted a user experiment with language experts to quantify gestures' properties. Next, we developed a gesture-based system that facilitates surgeons to control the medical-imaging software. Next, a usability study was conducted with neurosurgeons; standard usability metrics were measured, and their intercorrelation was studied.

Results: Statistical regression analysis showed that the v scores were significantly correlated with u scores ($R^2 \approx 0.4$, $p < 0.05$). Results show that there is a strong signal indicating that v and u are correlated. Hence, usability studies can be conducted with a relatively small number of surgeons, as u scores can be estimated from the v scores.

Keywords: *Gestures, Gesture Recognition, Agreement Analysis, Gesture Elicitation, Usability, Participatory Design.*

1. Purpose

Asepsis requirements are critical in the Operating Room (OR) and need to be fulfilled to avoid the spread of nosocomial infections and any risk of contamination (Spagnolo et al. 2013). This is especially timely given the current state of the COVID-19 pandemic. As such epidemic outbreaks occur, newer and alternative forms of interaction with health-related technologies will be required. In the OR, surgeons are required to navigate patients' medical images in order to recognize important anatomic landmarks and identify potential lesions in the brain (Wang et al. 2014; Sánchez-Margallo et al. 2017). However, current standard devices, such as the keyboard and mouse, pose a major drawback, as surgeons cannot have a direct contact with these nonaseptic devices. Hence, surgeons seek help from surgical assistants for manipulating the Picture Archiving and Communication System (PACS), which can cause miscommunication problems and lead to errors in the procedure (Hurstel and Bechmann 2019; O'Hara et al. 2014). In some cases, it is necessary for the surgeon to interrupt the intervention, directly manipulate the PACS to obtain the information needed, and go through the scrubbing and gloving process again. These solutions are known to cause significant delays in medical procedures (Sánchez-Margallo et al. 2017; Wipfli et al. 2016).

In this regard, touchless interfaces powered by gestures (Farhadi-Niaki et al. 2013) and speech have become an attractive solution, as they have allowed surgeons to control the PACS by themselves while maintaining sterility in the OR (Massaroni et al. 2018; Stuij 2013). In addition, these interfaces offer an intuitive and a natural way (Tomasello et al. 2019) of communicating with machines and smart devices, as they resemble human-human interaction (Wipfli et al. 2016; Rosa and Elizondo 2014). However, the choice of gesture languages (the gestures used to control these devices) plays a critical role in determining the usability and the acceptability of these interfaces. Hence, the principles of participatory design (e.g., gesture elicitation studies) are commonly used to involve end users at the early stages of the design process in order to gather information related to domain constraints and their preferences (Spinuzzi 2005; Muller and Kuhn 1993). The next step in the design process is to use agreement analysis based on majority voting to identify the best lexicon among a set of languages.

Nevertheless, there are several scenarios when agreement analysis fails to identify the best languages. For instance, when subjects do not agree on a gesture, agreement-based approaches are bound to select a random gesture from a given pool of gestures (Gonzalez et al. 2018). Next, these approaches are based on majority voting among surgeons and do not optimize for the task performance and user satisfaction (Vatavu and Wobbrock 2015). These limitations can be addressed by conducting usability studies to quantitatively identify the best languages based on how well surgeons are performing the image manipulation tasks (Madapana et al. 2019).

Furthermore, we proposed a gesture selection method, referred to as a VAC (Vocabulary Acceptability Criteria) study, as an alternative to the agreement analysis (Gonzalez et al. 2018). This method was mainly based on assessing the linguistic and cognitive properties of gestures with an end goal of identifying a best gesture language. In contrast to the usability studies, which are conducted with end users, VAC studies are conducted with speech and language professionals (SLPs). The main goal of this part of the work is to test our hypothesis that there is a correlation between gestures' linguistic properties and usability metrics. In this regard, the gesture lexicons and their corresponding VAC scores presented in our previous work (Madapana et al. 2018; Gonzalez et al. 2018) are used in this paper to validate our methodology.

To this end, we adopted a systematic approach to develop a gesture recognition system that allowed surgeons to manipulate the PACS in a touchless manner. The first step in the process was to identify

the typical commands in the PACS software that are routinely used in the OR to manipulate the MRI sequences. Next, we conducted a gesture elicitation study with surgeons to take note of the preferences of surgeons so that they can be incorporated into the gestural system. Next, we conducted a human factors study with speech and language experts at Purdue University to find and annotate the important qualitative aspects of gestures. Then, we conducted a usability study with surgeons to measure the task performance metrics (or usability metrics) in a quantitative manner. In this study, surgeons were asked to perform two image manipulation tasks using the gestural system, and usability metrics such as *quickness, learnability*, and *effectiveness* were measured. These usability metrics are inspired from the works of Farhadi et al. and Bhuiyan et al. (Farhadi-Niaki et al. 2013; Bhuiyan, Picking, and others 2011), who studied the usability of a gesture-based control of common desktop tasks. Using regression techniques, it was demonstrated that the usability metrics are correlated with the VAC ($R^2 \approx 0.4$, $p < 0.05$). Hence, we concluded that the obtained correlation coefficients can be used to predict the usability scores of new gesture lexicons, obviating the need for another usability study. The main objectives of this work are to 1. identify the typical commands in the PACS software; 2. obtain gesture vocabularies from domain experts (surgeons in this case); 3. determine the best gestures using qualitative studies (VAC: Vocabulary Acceptability Criteria) and usability studies; and 4. conduct usability study with surgeons to compare three interaction modalities: gestures, speech, and keyboard and mouse interfaces.

2. Scope

In the past 10 years, hand gesture-based interaction systems for PACS control have been introduced in the OR in order to reduce the risks of diseases spread through direct contact (O'Hara et al. 2014). Most of the research in this field focused on algorithms and sensors enhancement (Fukumoto, Suenaga, and Mase 1994; Vatavu 2012; Strickland et al. 2013; Jost et al. 2015). Such studies assumed a smaller and constrained set of commands (fewer than 20 gestures) for PACS operation in order to relax the complexity of the systems.

In order to maintain sterility in the OR, it is common practice for a surgeon to convey the image manipulation commands verbally to an assistant who is sitting near the computer and operating the PACS. However, studies show that this approach involves verbal miscommunications that lead to significant delays in the surgical procedures (Ebert et al. 2012; Johnson et al. 2011; Wipfli et al. 2016). Recently, gesture-based interaction was compared against direct manipulation with a mouse and against verbal transfer of information to an assistant (Wipfli et al. 2016). Their results showed that the gesture modality was significantly more efficient than verbalizing the instructions. Multimodal systems have also been explored in the area of automatic recognition in the OR. Some of these studies focused on interfaces that allow both voice and gestural commands (Mentis et al. 2015). Particularly, the work by Grange, Fong, and Baur (2004) aimed to design an architecture that would allow natural gestures only, leaving all other actions to voice recognition. The work in Lee et al. (2012) expanded the multimodal concept by adding an autonomous modality, in which the system determines the actions to assist the surgeon based on the contextual information. Alternatively, gesture design from the usability front was studied in Lee et al. (2012) and Norman (2010). For example, the work in Johnson et al. (2011) and O'Hara et al. (2014) explored the sociotechnical aspects that constrain a gestural interface in the OR. Additionally, Nacenta et al. (2013) studied the memorability of gestural interfaces by comparing random, predesigned, and user-defined lexicons. The results of this work showed that a user can recall 15-16 gestures with a very little learning time if they are customized.

Agreements studies are a very common first step when deciding on a lexicon for a gestural interface. Agreement analyses are also a common part of what is referred as elicitation studies (Kray et al. 2010; Vatavu 2012; Vatavu and Wobbrock 2016). The agreement found tends to vary greatly according

to the number of commands, the type of interface, the number of subjects, and the way that similar commands are grouped together (Vatavu 2012; Mauney et al. 2010). In addition, some studies have tried different grouping taxonomies to determine the properties in gestures that give a higher consensus. In Mauney et al. (2010), the gestures were divided into symbolic actions (0.35 agreement) and direct manipulation (0.18 agreement). In Luthra and Ghosh (2015), the gestures were classified as metaphorical, symbolic, physical, and abstract. Others (Kray et al. 2010; Wobbrock, Morris, and Wilson 2009) came up with a taxonomy to describe and classify their gestures and measure the consensus according to those properties. Other user consensus works are not limited to the 2D manipulation of the space; in work by Piumsomboon et al. (2013), an elicitation study for augmented reality tools was conducted, finding a 29% agreement between subjects.

The vast majority of elicitation studies use the metric provided by Wobbrock, Morris, and Wilson (2009) and Vatavu and Wobbrock (2016). The best gestures are the ones that are chosen by a majority of participants. However, when all participants choose a different gesture, the agreement rates tend to be very low, and agreement analysis fails to identify the best gestures. This is a common problem, especially in *unconstrained gesture elicitation* studies, in which subjects can propose a gesture of their choice without constraining themselves to a list of predetermined gestures. As a result, several research works explored gestures' qualitative properties in conjugation with agreement analysis to determine best gestures when agreement rates are low (Glowinski et al. 2011). As a part of this project, we conducted a user study with speech and language professionals (SLPs) to develop a set of qualitative characteristics referred to as Vocabulary Acceptability Criteria (VAC) (Gonzalez et al. 2018). In this study (known as the VAC study), the six properties proposed were *iconicity, simplicity, efficiency, compactness, saliency,* and *economy of movement*.

In addition to VAC-like studies, usability studies are often conducted with end users (surgeons in our case) to identify the best gesture languages (Piumsomboon et al. 2013). In the context of gesture design, the goal of usability studies is to identify the gesture lexicon that optimizes the task performance metrics of surgeons. These metrics include but are not limited to *task execution time, error rate*, and *ease of use* (Yen and Bakken 2012; Farhadi-Niaki et al. 2013; Bhuiyan, Picking, and others 2011). Farhadi et al. conducted a usability study to compare three different modalities for a 3D game: *haptic 3D mouse, static gestures*, and *dynamic gestures*; these were based on eight criteria (*precision, efficiency, ease of use, fun to use, fatigue, naturalness, mobility,* and *overall satisfaction*). Tsai et al. showed that children and adults outperformed the elderly in tasks related to smartphone manipulation with respect to the time taken to finish a task (Tsai, Tseng, and Chang 2017).

In the context of gestural interfaces, several usability studies were conducted in clinical settings to assess a set of gesture vocabularies (Opromolla et al. 2015; Soutschek et al. 2008). The quantitative usability metrics that are used in these studies include *system accuracy, memorability, learnability, intuitiveness,* and *task completion time* (Mewes et al. 2016). Furthermore, Ebert et al. conducted a user study to compare the usability of keyboard and mouse interfaces with respect to speech and gesture-based interfaces (Ebert et al. 2012). It was found that the *task completion time* was significantly higher for touchless interfaces. Overall, it was noticed that the number of subjects in these user studies varied from 10 to 15, as neurosurgeons are busy and constrained in the amount of time they can allocate.

In this context, our main objective was to develop a full-stack touchless interaction system, considering various aspects of gesture design and surgeon usability. In this process, we developed novel techniques to measure agreement analysis using gesture descriptors, and we established a correlation between gestures' qualitative properties (VAC) and the usability metrics. When such correlation existed, it was shown that the number of surgeons for the usability study can be significantly reduced. Last, the usability study

conducted with surgeons showed that they prefer gestures and speech-based systems, as they are more natural and intuitive to use; however, they are considered to have a learning curve, which led to increased task completion times.

3. Methods and Results

Our methodology has been subdivided into five studies (Task A1, Task A2, Study 1, Study 2, and Study 3). These subtasks are described in more detail here.

3.1 Task A1: Finding Typical Functions of PACS

Objective: The goal of this study was to obtain the typical functions of the Picture Archiving Communication Systems (PACS). Synapse, a popularly used PACS radiology image browser, was used in this project. Three medical image manipulation tasks, encompassing most of the functionalities of Synapse (the PACS system utilized), were considered.

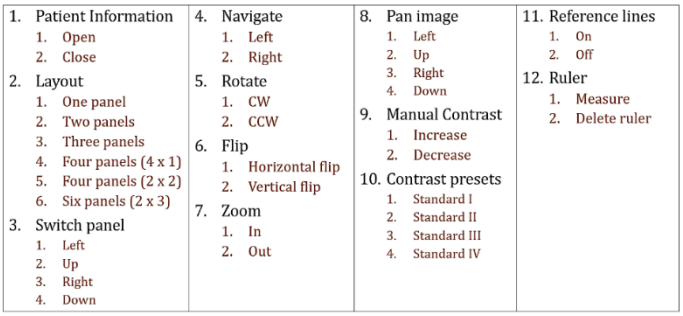

Figure 1. Typical commands of PACS.

Participants: This study consisted of nine neurosurgeons operating the PACS software (Synapse). This study was approved by the Institutional Review Board (IRB) of the Indiana University School of Medicine. Signed informed consents were obtained from the participants prior to beginning the study.

Experimental Protocol: Initially, each neurosurgeon was asked to accomplish these tasks using keyboard and mouse interfaces. By tracking the menu choices, mouse motions, and selections, all functions required to complete the tasks using the PACS system were collected. The outcome of this study is an extensive list of 34 PACS commands, as depicted in Figure 1.

3.2 Task A2: Determining Surgeons' Gestures for PACS Commands

Objective: The goal of this task was to conduct a gesture elicitation study with surgeons to identify the gestural preferences of the surgeons, as they have the domain knowledge about the constraints in the OR.

Participants: This study consisted of nine neurosurgeons eliciting gestures for the PACS software (Synapse). This study was approved by the Institutional Review Board (IRB) of the Indiana University School of Medicine. Signed informed consents were obtained from the participants prior to beginning the study.

Experimental Protocol: Initially, the subjects were asked to follow three predetermined steps in the same order: 1) gesture design on a drawing sheet: required subjects to design and draw the gestures (as shown Figure 2) corresponding to each of the 34 commands of Synapse on a drawing sheet; 2) gesture illustration: required the subjects to perform each of the chosen gestures in front of Microsoft Kinect v2; 3) manipulation task using Wizard-of-Oz setup: required the subjects to perform a tumor identification task using the chosen gestures, following a Wizard-of-Oz experimental setup. In this way, we obtained the gestural preferences of the neurosurgeons for the 34 PACS commands. These gestures will be further used in the next experiments to study the agreement among subjects, to provide guidelines for gesture selection, and to determine the Vocabulary Acceptability Criteria (VAC) to quantitatively evaluate gestures.

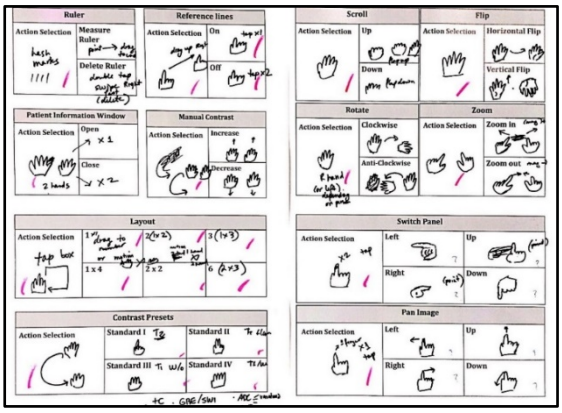

Figure 2. Gestures drawn by the surgeons on the drawing template.

This design allowed the subjects to naturally control the software using gestures without realizing that the investigator (the wizard) interprets their gestures and controls the program. This experimental design isolates the subject from the wizard to capture their natural movements. This study resulted in a total of 306 gestures (including combined and single gestures). In other words, we obtained nine lexicons corresponding to nine surgeons, each containing a total of 34 gestures (34 commands). These gesture lexicons were further utilized to analyze the how well surgeons agreed on gestures and/or its properties. Subsequent data analysis led to novel techniques related to agreement analyses and design guidelines for gesture-based interfaces. The proposed methodology and the results associated with these findings are discussed in detail in the next section.

3.2.1 Agreement Analysis

Agreement analysis quantifies the degree of preference among the users. It is a well-known fact that the participatory designs consisting of user elicitation studies are crucial for developing effective and usable interfaces. This analysis is especially beneficial in expert domains (i.e., neurosurgeons, urologist, radiologists), because these populations have intrinsic knowledge about the environment that shapes the gestures that they elicit. Thus, determining the gestures without the participation of end users in the early stages of a design process can potentially lead to suboptimal lexicons and unusable interfaces. The gestures obtained in task A2 were studied to assess the level of agreement among neurosurgeons. This consensus was obtained through two methods: Method 1. the state-of-the-art metric used for gestural agreement in a group; and Method 2. a metric based on a novel representation for gestures that uses their semantic descriptors. The following sections explain both metrics in detail.

Method 1

The state-of-the-art approach proposed by Wobbrock et al. (2008) is widely used in the literature to measure the level of consensus among subjects. That method has two stages: grouping and evaluation. In the grouping stage, the gestures for the same command are clustered by similarity. Thus, all the gestures that are placed in the same group are considered equal. Then, the consensus for each command is obtained by summing the squares of the size of each group, divided by the total number of gestures available for that command. This metric takes a value of one when there is complete agreement (consensus) (i.e., all the

gestures chosen for a command were identical). When there is no agreement, the value of the metric is 1 over the number of gestures for that command.

Let us first define the notations that will be used throughout the paper. Let C be the total number of commands or referents for a system. Let P_r be the set of gestures elicited by the user for the r^{th} command, in which $r = 1, \dots, C$. Additionally, let P^i_r be a subset of gestures for the r^{th} command that are considered identical. Thus, $|P^i_r|$ would be the number of identical gestures in the i^{th} set for the r^{th} command. Finally, let $u_r < |P_r|$ be the number of unique gestures for the command r. The total number of gesture examples (N_r) for the command r can be represented similarly.

The agreement index proposed by Wobbrock et al. (2009) is currently the most commonly used metric, and it is defined as follows, in which A_r is the level of agreement for the r^{th} command.

$$N_r = \sum_{i=1}^{u_r} |P_r^i| = |P_r| \;;\; A_r = \sum_{P_i^r \subseteq P_r} \left(\frac{|P_r^i|}{|P_r|}\right)^2 \;;\; A_{overall} = \frac{1}{C}\sum_{r=1}^{C} \sum_{P_i^r \subseteq P_r} \left(\frac{|P_r^i|}{|P_r|}\right)^2$$

There are two major problems with this approach. The first one is that, when there is no agreement at all, the agreement value is not zero. The second one is that the literature does not provide good qualitative interpretation of this metric. For example, if an agreement of 20% is found, this does not mean that 20% of the group agrees on a gesture. Thus, this metric is of little use when looking for an optimal lexicon to control a medical software. To compensate for the limitations of the state-of-the-art approach, a novel method to measure the level of agreement is proposed and explained in the next section.

Method 2

Previous efforts to find the level of agreement considered gesture as a concrete entity, ignoring embedded properties of the gestures (e.g., hand shape, hand motion trajectory, plane of motion, etc.). We propose to utilize the same set of semantic descriptors used for the *Heuristic Generation* (Section 2.2.1) to compute the level of agreement for each command.

To measure the similarity between two gestures, the well-known Jaccard metric J was used. This metric is a suitable method to evaluate the distance between two sparse binary vectors. The overall agreement with Jaccard can be calculated using the equation below. This formula averages the agreement between all the possible pairs for the same gesture, for all 34 gestures. In this equation represents the total number of commands, and $_g$ represents the total number of gestures per command. S and ^{k} represent the gesture examples i and j for the command k.

$$A_{overall} = \frac{2}{N_c\, N_g\, (N_g - 1)} \sum_{k=1}^{N_c} \sum_{j=i+1}^{N_g} \sum_{i=1}^{N_g} J(s_i^k, s_j^k)$$

The square root of $A_{overall}$ represents the percentage of subjects that agreed on a gesture (N_{eq}). This offers a very clear and intuitive interpretation for the consensus between surgeons. The next section shows the results of the agreement analysis for Metric I and Metric II. Additionally, agreement is also studied for the full gesture (context and modifier together), the context alone, and the modifier alone.

$$N_{eq} = \sqrt{\frac{2}{N_c\, N_g\, (N_g - 1)} \sum_{k=1}^{N_c} \sum_{j=i+1}^{N_g} \sum_{i=1}^{N_g} J(s_i^k, s_j^k)}$$

Experiments and Results

Splitting commands into context and modifier: The commands obtained in Task A1 were grouped into meaningful clusters based on semantic and numerical similarity, as shown in Figure 1. This allowed us to reduce the number of gestures that surgeons would need to remember. We adopted a context-based approach, in which the same gesture would mean different things based on the context (the layer) used. This gesture-based interface's architecture consists of two layers: context and modifier. The context represents the general action that the surgeon wants to perform (i.e., Zoom, Pan, Scroll), and the modifier represents the different options for that command (i.e. In and Out for Zoom or Up, Down, Left, and Right for Pan and Switch Panel). This approach resembles interpersonal communication, in which limited numbers of gestures are reutilized and their meaning is understood according to the current context.

Experiments: We performed an agreement analysis using both the traditional method (Metric I) and our approach based on semantic descriptors (Metric II). The results are shown in Table 4. This table shows the agreement index for the full gesture (context + modifier), context only, and modifier only. The reason for splitting the results into these three categories relies on the fact that the command's context tends to be more abstract (i.e., open PIW, change layout, select contrast present, etc.), whereas the modifier actions tend to correspond to a more direct effect on the image (i.e., up, down, left, right, zoom in, zoom out). Thus, the gestures for the contexts had a very high variance between subjects when compared with the gestures for the modifiers (refer to Table 1).

Results: The highest agreement index was found with the modifier gestures, with a value of 0.34 for Metric II (see Table 4). A one-tailed, paired, sampled T test ($= 34 - 1 = 33, \alpha = 0.05$) was performed to ensure that the difference between the metrics was statistically significant.

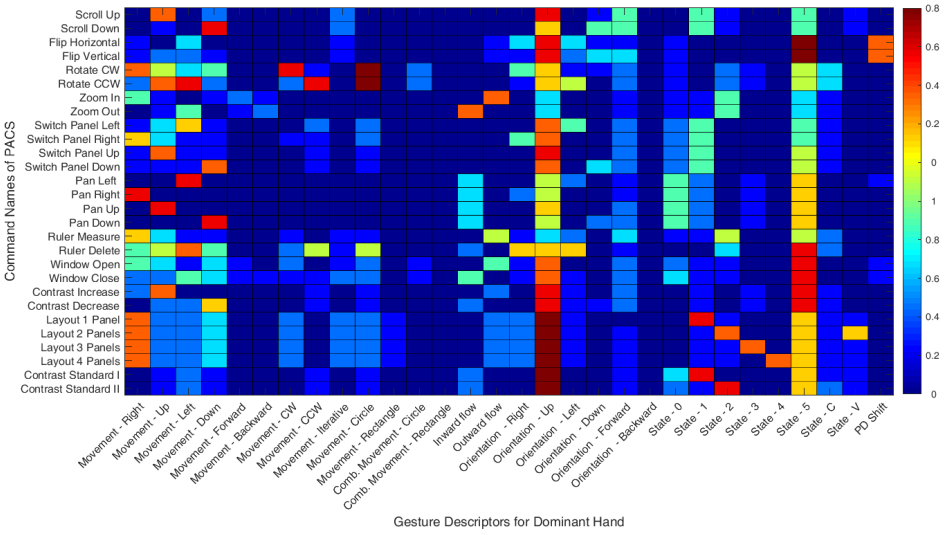

Figure 3. Visualization of popularity of descriptors for each command across all subjects. The red colors are closer to 1 and the blue color is closer to zero.

The null hypothesis stated that the agreement indices for Metric I and Metric II were the same, while the alternative hypothesis stated that the index of Metric II was greater than the index of Metric I. The difference between methods was statistically significant ($p < 0.05$) for the modifier and the gesture + modifier. The context did not show a significant difference between the metrics, because it was too low in both cases.

These results show that more abstract commands tend to produce a lower agreement between subjects (this is an intuitive finding). Additionally, Metric II captures a greater agreement index, because the subjects are choosing gestures that have common properties even when they are different. In addition, the results contradict the intuition that the agreement (without semantic descriptors) between subjects in the same domain should be higher than 30% agreement for 70% of the subjects (Stern, Wachs, and Edan 2008). This

could be explained by the fact that many PACS commands are complex hard to translate to gestures that are highly iconic, even when working solely with neurosurgeons (refer to Figure 3).

Interpretation of agreement: The square root of Metric II can be directly interpreted as the percentage of subjects that agreed on one gesture. This means that, on average, 58% of the subjects agreed on a gesture for the modifier of each command (refer to Table 1 and Figure 4). Again, the modifier is the part of the gesture that shows the highest level of agreement. All the top commands involve gesturing a number in the modifier. This means the surgeons have a very standardized way of gesturing the numbers from one to four.

Table 1. Consensus measured by Metric I (State of the Art) and Metric II (The Jaccard distance using semantic descriptors). (= statistically significant, $p < 0.05$)*

Category	***Metric I (SOA)***	***Metric II (SDs)***
Context	0.13 ± 0.02	0.1146 ± 0.09
Modifier	0.23 ± 0.13	**0.34 ± 0.14***
Context + Modifier	0.13 ± 0.02	0.29 ± 0.06*

3.3 Study 1: Develop Vocabulary Acceptability Criteria for Gestures

The goal of this study was to develop Vocabulary Acceptability Criteria (VACs) that can potentially explain the qualitative aspects of the gestures. Furthermore, the obtained gestures were compared, ranked, and evaluated using these VACs in order to obtain a best gesture lexicon. The final gesture recognition interface consists of these carefully selected gestures so as to enhance the usability of the overall system. This study consists of three major parts: Part I – Creation, Part II – Evaluation, and Part III – Validation.

3.3.1 Part I: Creation of Vocabulary Acceptability Criteria (VACs)

The objective of this part was to come up with a series of criteria to evaluate gesture lexicons. We named these criteria VAC, which are means to quantify, compare, and rank the gestures with an end goal of obtaining the best lexicons from an available set. For obtaining VACs, a discussion panel composed of 17 SLP experts was recruited through the Speech and Language Department of Purdue University. The reason behind choosing the SLP experts was that, in literature, it has been shown that speech and language expertise could help design more effective and user-friendly interfaces.

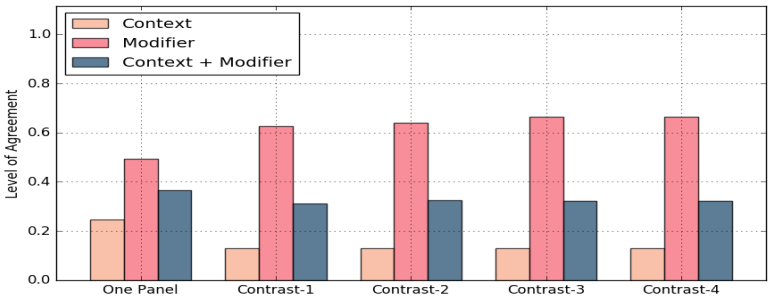

Figure 4. Consensus of the top five gestures with highest level of agreement.

We used an adapted version of the Delphi method to engage experts in discussion. The Delphi method is a structured communication technique to systematically engage a panel of experts in discussion. In this method, participants are encouraged to discuss their viewpoints on the topic being discussed and to present their opinions and reason behind them. Furthermore, participants are encouraged to revise their original opinions in light of opinions of other participants. This process is supposed to decrease the variation in opinions/choices with time and converge to better/ correct choices.

This part of VAC creation lasted for about an hour and a half. First, the experts were shown two practice gesture lexicons from the ChaLearn Gesture Dataset (CGD 2011), which is a standard dataset in gesture recognition research. Then, they were shown a random sample containing 3% of the gestures from the lexicons collected in task A1. The purpose of showing such a small percentage was to prevent a bias toward the surgical lexicon. After every three gestures, the experts were asked to come up with generalizable and possibly independent (orthogonal) gesture attributes (VACs). After two rounds of discussion, the experts proposed 18 VACs, namely *contrast, memorability, replicability, consistency, intuitiveness, combinatorial, distinct hand shape, repetitiveness, compactness, location on body, iconic, emulation, distinctiveness, complexity, directionality, efficiency, multistep,* and *visual saliency*. Then, the experts were asked to merge the VACs that represented the similar attributes (similar meanings). VAC is in a three-dimensional space spanned by time, meaning, and the space. The merge resulted in the following set of six final VACs: 1. ***Iconicity*** - how much gesture looks like the command that is associated with; 2. ***Simplicity*** - how straightforward are the movements; 3. ***Efficiency*** - capability of conveying more information in less movement; 4. ***Compactness*** - how much the gesture covers the space around the body frame; 5. ***Salience*** - how discriminative is the movement of the hand; and 6. ***Economy of Movement*** - amount of movement involved in the gesture.

3.3.2 Part II: Evaluation of Gestures using the Created VACs

The objective of the evaluation step was to assign six VAC scores (one for each VAC) to each of the 252 gestures. For this step, an additional group of 12 SLP grad students was added to the existing group of 17 SLP experts. Each expert was assigned one command (i.e., nine gestures corresponding to that command). However, each student was assigned two commands to evaluate. For this study, the number of commands was reduced from 34 to 28. The reduced list of commands is shown in Figure 1. The commands, such as Layout four and six panels, that are not often used by surgeons were ignored in order to increase the number of replications. Overall, each expert took about 45 minutes, whereas students took about an hour and a half to accomplish the assigned task.

A block randomization strategy was followed to assign commands to experts and students. As discussed in the previous report, the commands were organized as context and modifier. For example, Pan is referred as context, whereas pan up, down, left, and right were considered its modifiers. Hence, the commands can be classified into 12 command groups. In the above example, Pan is considered as a group. Similarly zoom, manual contrast, etc. are command groups. All the commands (both context and modifiers) are completely randomized in the beginning of the experiment. Experts were given the preference, and a modifier from each group was assigned to each expert in a sequential manner.

ABC	ADE	AFG	AHI
	BDE	BFG	BHI
	CDE	CFG	CHI
		DFG	DHI
		EFG	EHI
			FHI
			GHI

Figure 5. A minimal set of three-way combinations necessary to establish ranking for nine elements: A, B, C, D, E, F, G, H, and I. To rank nine elements requires 16 comparisons.

Once one modifier from all groups was assigned to the experts, the next modifier was assigned to the rest of the experts until all the modifiers were assigned to one of the subjects. Students were assigned the commands in an equivalent manner, starting from the commands remaining after completing the assignment to the experts. However, it was ensured that each of the students was assigned two commands, whereas experts were assigned only one command. Hence, 13 $(17 + [12 \times 2] - 28)$ of the 28 commands had two replications.

A variant of the pairwise comparison method was used to obtain the VAC scores. For each command and a VAC, a set of three-way comparisons among the nine respective gestures was presented. The expert had to order these three gestures according to the VAC criteria that was shown (from low to high). Figure 5 shows the minimum set of three-way comparisons necessary to infer a full order between nine gestures.

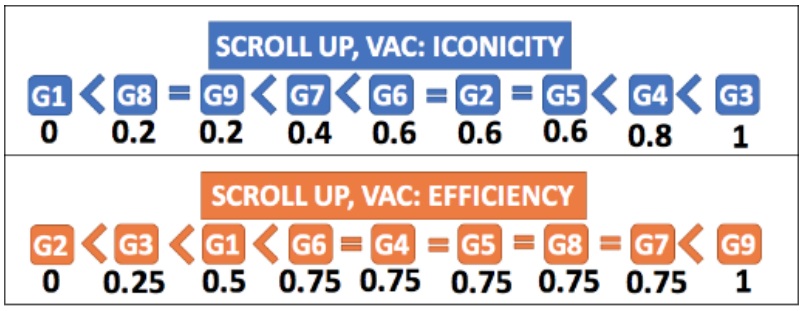

Figure 6. Two examples of score assignment for the gesture scroll up. Iconicity is illustrated on the top and Efficiency on the bottom. The ranking assigned by the ordering is mapped to a score between 0 and 1.

The sequence in which the comparisons were presented was completely randomized. A VAC score between 0 and 1 was inferred from the ordering generated by the SLPs, as illustrated in Figure 6. In the case of repetitions, the overall score was computed as an average of the scores obtained from multiple SLPs. Through this process, a total of six scores (one for each VAC) was assigned to each individual gesture, generating a total of 1512 ($28 \times 9 \times 6$) evaluations. The scores obtained for each command in Figure 1 across each lexicon and VAC are shown in Tables from 5 to 10.

By taking an average of the VAC scores across all the 28 commands, we found that not all of the lexicons were equally good. A pairwise T test was performed between the lexicons that had the top scores (namely, 8, 6, and 1) and the rest of the vocabularies. A statistically significant difference was found. Thus, we can claim that, according to the developed Vocabulary Acceptability Criteria, lexicons 8, 6, and 1 were better than the rest, as depicted in Figure 7. We replaced the VACs of *Complexity* and *Amount of movement* by *Simplicity* and *Economy of movement*, respectively. Because the VACs of *Complexity* and *Amount of movement* indicate a better performance when the value is close to zero, we used a complementary complex so that all VACs would indicate "good performance" when their score is close to 1.

	VACS						
Lexicons	Iconicity	Simplicity	Efficiency	Compactness	Salience	Economy of movemement	
1	0.43	0.61	0.63	0.6	0.57	0.69	
2	0.44	0.37	0.37	0.45	0.4	0.44	
3	0.52	0.28	0.24	0.25	0.41	0.24	Bad Lexicon
4	0.5	0.51	0.48	0.52	0.48	0.49	
5	0.5	0.45	0.54	0.53	0.54	0.54	
6	0.64	0.75	0.79	0.53	0.65	0.66	Good Lexicon
7	0.49	0.46	0.46	0.42	0.36	0.38	
8	0.52	0.7	0.69	0.74	0.68	0.75	
9	0.61	0.48	0.46	0.4	0.64	0.39	

Figure 7. Mean VAC scores of lexicons.

3.4 Study 2: Studying the Relationship between VACs and Usability Metrics

The goal of this study is to measure the usability metrics of gesture lexicons generated from VACs and find out the correlation between VAC scores and usability measures. Usability metrics are means to evaluate gesture lexicons obtained from the gesture elicitation study. The hypothesis is that lexicons corresponding to higher VAC values are more usable (i.e., there is a positive correlation between the VAC values and usability metrics). There are six major usability metrics, as defined here. The IRB, statistical design, and development of gesture recognition interface for this study are partially completed and still in progress.

1. ***Task Completion Rate:*** It is the ratio of tasks completed to the total tasks. A task is considered complete when a user finishes the task from beginning to end without any critical errors.
2. ***Time of Task:*** It is the time required to complete a task.
3. ***Error-Free Rate:*** It is the proportion of the participants who finished the task without any error.
4. ***Critical Errors:*** Critical errors are the errors that either lead to wrong outcome or do not lead to completion of task.
5. ***Non-Critical Errors:*** These errors are in general recovered by the subjects and don't lead to unexpected outcomes, though they increase the time on task and decrease efficiency.
6. ***Learnability and Memorability:*** These parameters are used to evaluate the accuracy and time taken by subjects to perform gestures during the task.

Population: Twelve medical professionals, including nurse practitioners and neurosurgeons with more than 2 years of experience in operating PACS software known as Synapse, were recruited on a voluntary basis. The subject pool consisted of eight men (age: 36 ± 7) and four women (age: 38 ± 7). Overall, there were seven neurosurgeons and five nurse practitioners in the subject pool. All were assigned to the same pool regardless of their position, as they were equally experienced with the Synapse software. This study was conducted at Goodman Campbell Brain and Spine Center, Indianapolis, which is a part of the Indiana University School of Medicine (IUSM). Written informed consent was obtained from all subjects.

Experimental Protocol: Let us start by defining notations. Let L1, L2, … LR be R gesture lexicons. Let there be S participants and T tasks in the usability study. Let v1, v2, ... vM be the M VAC and u1, u2, ... uN be the N usability metrics. We conducted two pilot studies and noticed that performing each task using our gesture recognition system takes approximately 15 minutes. Given that surgeons are time constrained and that the time allotted for this experiment was 1 hour, we created two possible designs for the task at hand: 1. Each subject performs one task using all lexicons; and 2. Each subject performs all the tasks using a single lexicon. We followed a completely randomized procedure to determine the order of task and lexicon assignment. For instance, if there are two lexicons, two tasks, and two subjects, then subject 1 is assigned <L2, T2, T1> and subject 2 is assigned <L1, T1, T2>. As a part of the experimental design, subjects were first asked to go through a training procedure to help them get familiarized with the task and the gesture recognition system.

Gesture Recognition Interface: The first step in the pipeline was to develop a gesture recognition interface that allows surgeons to manipulate the PACS in a touchless manner. This system was equipped with fail-safe mechanisms to cope with the errors associated with the gesture performance and the recognition algorithms. Figure 8 shows a surgeon controlling the MRI software using his hand gestures.

A Microsoft Kinect camera was used to record the RGB-D information of the subjects. This information was processed to obtain 3D body skeleton features from depth data. Next, a Deep Learning Neural Network was used to create the 2D hand skeleton features from the RGB videos using Convolutional Pose Machines (Pavllo et al. 2019). An ensemble of Support Vector Machines (SVM) classifiers was trained based on body skeleton features, hand skeleton features, and a conjugation of body and hand features.

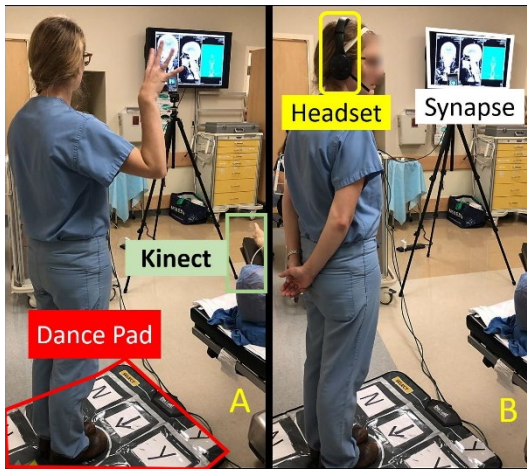

Figure 8. Surgeon controlling MRI software using gestures and speech.

These algorithms were trained on a per-lexicon basis so that there were considerably fewer target classes, thereby achieving higher accuracies of recognition. Hence, there were four trained models, one for each gesture lexicon. The probabilistic predictions obtained from this ensemble of models were combined to make the final prediction. Figure 9 depicts the flowchart of the gesture recognition pipeline. It was found that the final gesture recognition accuracies varied from 80% to 95% depending on the lexicon.

Though the gesture recognition accuracies are relatively high, they are not enough for a real-time gesture recognition system. In other words, surgeons may get frustrated when performing gestures that are hard to recognize (low accuracies), as they need to repeat them multiple times before it is accurately recognized. In this regard, we developed a feedback and a fail-safe mechanism in order to make the system robust to errors. This mechanism is facilitated by a dance pad that allows surgeons to visually navigate across the top predictions of the system and make a final decision. In other words, it further allows them to navigate through the top five predictions, with all existing commands if necessary, before selecting the final command using the dance pad.

This feedback mechanism acts as an acknowledgement system, as the surgeon gets a chance to accept/reject the command after the gesture is performed. The top five command predictions were shown to the users, and they got an option to pick one of those options or utilize another option that will let them look at all the commands on the screen. The last option is a fail-safe mechanism in the sense that users would be able to finish the task independently without any human intervention. Once the gestures were recognized and mapped to one of the commands available in the Synapse software, we developed an automation software to automatically execute the commands.

Usability Metrics and Annotations: Three usability metrics were considered to identify the best gesture lexicon among a group of lexicons. The first characteristic, *quickness*, was defined as an average rate at which a gesture was performed by the participants. It was measured as an inverse of the time taken to perform a gesture. The second property, *learnability*, was defined as the ease at which participants learn to perform and recall the gestures in the lexicon. It was measured as the inverse of the frequency of focus shifts. The reason for this is that, when the gestures were committed to memory, there was no need to look at the cheat sheet with the commands and the pictorial representation of the gestures (thus reducing the focus shifts). The third entity, *effectiveness*, was defined as a measure of success in gesture performance. It was measured as the inverse of the frequency of errors committed while performing the image manipulation task using a particular gesture lexicon.

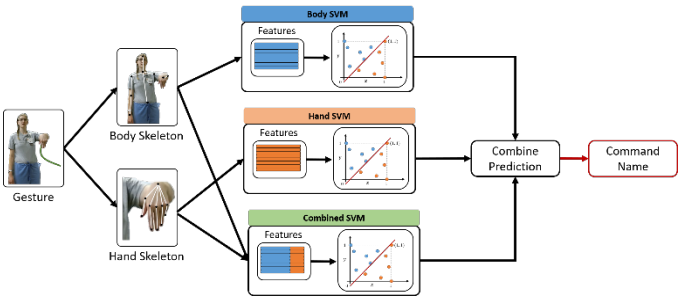

Figure 9. Gesture recognition pipeline.

Let us define some notations. Let the subscript *i* refer to the command i and superscript *j* refer to the gesture j and there are *G* gestures in total for each command. Let t_i^1, t_i^2, t_i^G be the time taken to perform the respective gestures. If a gesture is performed multiple times by the same or different subjects, let t_i^j define the average time taken to perform gesture *j* of command *i*. Let t_i^{min} and t_i^{max} be the minimum and the maximum time taken by all the gestures corresponding to the command *i* (i.e., $t_i^{min} = \min[t_i^1, t_i^2, \ldots t_i^G]$ and $t_i^{max} = \max[t_i^1, t_i^2, t_i^G]$). Let z_i^j, f_i^j, and m_i^j be the average number of occurrences, the average number of focus shifts and the average number of errors occurred, respectively. Let q_i^j, l_i^j, and e_i^j refer to the usability metrics of *quickness*, *learnability*, and *effectiveness*, respectively.

Now, they are mathematically expressed. Note that the usability metrics were appropriately normalized to ensure that they range from [0 to 1], and the gestures with higher usability indices were considered as better gestures. For instance, a value of $q^j_i = 0$ indicates that this particular gesture takes much longer to perform than $q^j_i = 1$. Similar interpretation is valid for other metrics such as l^j_i and e^j_i.

Next, the data obtained from the usability study were annotated with respect to the aforementioned metrics. In this regard, a software interface was developed to annotate user focus shifts and errors. The metric *quickness* was computed directly using the timestamps. Note that the time taken to navigate the acknowledgement dance pad was not considered, as we were only interested in the time taken to perform the gesture. Next, the metric *learnability* was measured indirectly using the focus shifts. To compute this metric, we first annotated if there was any shift of focus between the PACS screen and the drawing template 1 or 2 seconds prior to performing the gesture. The information related to the focus shifts was annotated by looking at the RGB video that was recorded by the Kinect camera. Similar to focus shifts, the metric *effectiveness* was measured using the frequency of errors. Overall, each gesture was represented as a three-dimensional vector, as there are three usability metrics.

Correlation Analysis: Let there be K gestures in total, M VAC and N usability metrics. Let $v_i \in [0, 1]^K \; \forall \; i = 1, 2, \ldots, M$ be the VAC vectors and $u_j \in [0, 1]^K \; \forall \; j = 1, 2, \ldots, N$ be the usability vectors for all gestures. Let $v = [v_1, v_2, \ldots, v_M]$ be the VAC matrix of dimension $K \times M$ and $u = [u_1, u_2, \ldots, u_N]$ be the usability matrix of size $K \times N$. Now, the goal was to study the multivariate correlation between both the quantities (u and v). Formally, we want to estimate the function $f([0, 1]^M \rightarrow [0, 1]^N)$: $V \rightarrow U$, in which V and U represent the VAC space and the usability space, respectively.
The correlation analysis was conducted at two stages: 1. univariate versus multivariate, and 2. overall and per-command scenarios. In the univariate analysis, the correlation between every VAC with respect to every other usability metric was considered. Hence, there will be $M \times N$ correlation tests. Note that g is a function that can be either linear or nonlinear.

$$u_j = g(v_i) \; \forall \; i \in \{1, \ldots, M\} \text{ and } j \in \{1, \ldots, N\}$$

However, in the multivariate analysis, each usability metric was correlated with a group of M VAC, which captures the interdependencies between the VACs.

$$u_j = g(v_1, v_2, \ldots, v_M) \; j \in \{1, \ldots, N\}$$

Next, the correlation analysis was subdivided into *overall* and *per-command* scenarios. In the *overall* scenario, intercommand differences were ignored, and the gestures were evaluated independently. Hence, all the gestures were considered simultaneously for the correlation analysis. However, in the *per-command* scenario, we wanted to study the intercommand differences and study each command separately. Hence, only those gestures that corresponded to a particular command were considered for studying the correlation.

	L1			*L2*			*L3*			*L4*		
Command Name	u_1	u_2	u_3	u_1	u_2	u_3	u_1	u_2	u_3	u_1	u_2	u_3
Scroll Up	0.8	0.9	0.9	0.0	0.0	0.6	1.0	1.0	0.0	0.6	0.9	1.0
Flip Horizontal	1.0	0.8	1.0	0.0	0.0	0.1	0.0	0.5	0.0	0.7	1.0	1.0
Rotate CW	1.0	1.0	0.7	0.0	0.0	0.4	0.6	1.0	1.0	0.6	0.8	0.0
Zoom In	0.9	0.9	1.0	0.0	0.0	0.7	1.0	0.8	0.0	0.6	1.0	0.9
Switch Panel Left	0.4	0.6	0.0	0.0	0.0	0.4	1.0	1.0	1.0	0.6	0.8	0.9
Pan Down	**0.0**	0.0	0.5	**0.0**	0.0	0.0	**1.0**	1.0	0.4	**0.7**	0.9	1.0
Manual Contrast Increase	0.2	0.1	0.4	0.0	0.0	0.7	1.0	0.7	0.0	0.1	1.0	1.0
Layout Two Panels	0.0	0.0	1.0	0.9	0.0	0.0	0.7	1.0	0.9	1.0	0.0	0.8
Contrast Presets II	0.8	0.2	1.0	0.4	0.0	1.0	1.0	1.0	0.0	0.0	0.4	0.8

Figure 10. Usability indices of gestures.

Experiments and Results: The raw data collected from this study was used to compute three usability metrics: *quickness* (u_1), *learnability* (u_2), and *effectiveness* (u_3). Figure 10 depicts the usability indices for

each of the 20 commands. Note that a value of zero indicates a bad gesture, and a value of one indicates a good gesture. Figure 10 shows only one command from each group. For instance, *scroll up* and *scroll down* belong to the same group called *scroll* and have approximately the same usability indices, as they are complementary gestures.

Question	Score (mean ± SD)	Min. Score	Max. Score
Successfully completed the procedure	3.2 ± 0.5	2	**4**
System was easy to work with	3.2 ± 0.7	2	**4**
Reduced the time taken to complete a task	1.6 ± 1.5	**0**	3
System did not generate frustration	2.1 ± 0.9	1	3
System was intuitive and easy to remember	2.3 ± 0.8	1	3

Figure 11. Descriptive statistics of user satisfaction for the gesture recognition system. Note that 0 and 4 were the minimum and maximum possible scores, respectively.

In this Table, u_1, u_2, and u_3 refer to the three usability metrices, and the first column refers to the commands in the Synapse software. The lexicons L1 and L2 were worst lexicons, and L3 and L4 were best lexicons, according to VAC. Note that there might be few good gestures in the worst lexicons and vice versa in the best lexicons. Hence, the usability indices of some gestures in the worst lexicons can be considerably higher. Given our hypothesis that usability metrics and VACs are proportional, we expect the usability indices of worst lexicons to be relatively lower than their best counterparts.

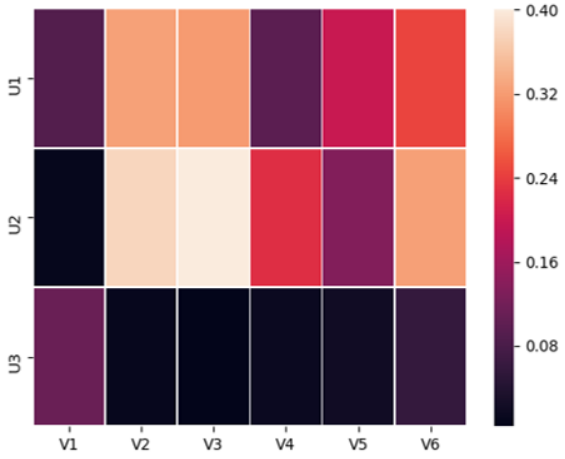

Figure 12. Univariate correlation between VAC and usability indices.

For instance, considering the gesture corresponding to *pan down* command, the value of u_1 was 0.0 for lexicons L1 and L2; however, it was 1.0 and 0.7 for L3 and L4, respectively. This trend of having low usability indices for the worst lexicons was observed for other commands, such as manual contrast, layout, rotate, zoom, and switch panel. However, there are some exceptions, such as the flip command (i.e., the value of u_1 was 1.0 for L1, 0.0 for L2 and L3, and 0.7 for L4). This indicates that the gesture for the flip command has relatively high quickness property for L1 and L4.

Furthermore, the usability indices for lexicon L2 were close to zero for all commands, indicating that this is a bad lexicon according to both the VAC and the usability metrics. Similarly, the usability indices for lexicon L3 and L4 were close to 1.0, indicating that most of the gestures in L3 and L4 were good according to VAC and usability metrics. However, for lexicon L1, half of the usability indices were less than 0.5, indicating that approximately half of the gestures were good and the other half were bad.

Multivariate Correlation: Once the usability indices were computed, our next goal was to study the dependencies between the VACs and usability metrics. In other words, our objective was to test the hypothesis that there was a direct relation between the VACs and usability metrics. Identifying such correlations would allow us to predict the usability indices given the VAC values. In this regard, the correlation analysis was conducted in two conditions: 1. overall conditions; and 2. command-dependent conditions.

In the command-dependent scenario, the gestures corresponding to each command were analyzed in isolation. For instance, in our case, there were four gestures for each command. Hence, the correlation

between the VAC and usability metrics was studied for those four gestures. However, in overall scenario, the differences between the commands was ignored (i.e., the correlation analysis was conducted for 80 gestures). Although the former approach considers the intrinsic differences between the commands and treats them independently, the latter approach ignores such differences. Note that the latter approach produces one correlation test per usability metric; however, the former approach produces 20 correlation tests (there are 20 commands) per usability metric.

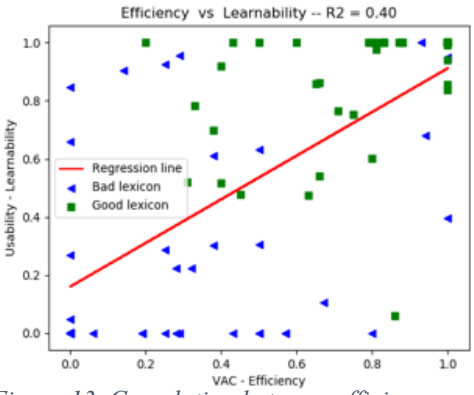

Figure 13. Correlation between efficiency and learnability.

In both of these conditions, we conducted univariate (refer to Figure 12) and multivariate analyses. In univariate analysis, each usability metric was correlated with each of the VACs separately, resulting in 18 (6 x 3) correlation tests. The correlation can be easily visualized in the univariate scenario, as there is only one dependent variable (VAC) and one independent variable (usability index). However, in multivariate analysis, each usability metric was considered to be a function of all VACs, hence producing three correlation tests (there are three usability metrics). Furthermore, the strength of correlation was reported using the coefficient of determination (R2) and *p* value.

Figure 13 depicts the correlation between the usability metric, learnability, and the VAC, simplicity. Note that the gestures corresponding to the worst lexicons were represented as blue triangles; the gesture in best lexicons were represented using green squares. Furthermore, the worst gestures were concentrated at the bottom left corner, whereas the best gestures were located at the top right corner of the plot.

Descriptive Statistics	Order of modalities
Which modality is easy to work with?	$S < G < \mathbf{K}$
Contributed to reduce the time taken to complete a task	$S < G < \mathbf{K}$
Contributed to reduce the frustration	$S = G < \mathbf{K}$
Natural and intuitive mode of operation	$K < \mathbf{S} = \mathbf{G}$
Overall, which modality do you prefer?	$S = G < \mathbf{K}$

Figure 14. Descriptive statistics of the user study to compare modalities.

At the end of the user study, subjects were asked to populate a satisfaction questionnaire that was an adapted version of the NASA-TLX. The score range for each question was 0-4, in which 0 represents strongly disagree and 4 represents strongly agree. Overall, it was found that 11 of 12 surgeons agreed that the gesture interaction system provided enough capabilities to successfully complete the image manipulation task. Similarly, 10 of 12 surgeons agreed that the system was easy to work with. However, six of 12 surgeons noted that gesture interaction system increased the amount of time taken to complete the procedure. The ratings obtained from the satisfaction questionnaire were summarized in Figure 11. On an average, surgeons provided a score of 3.2/5 for successfully completing the procedure and ease-of-use of the system. However, scores of 1.6 and 2.3 were given to reduction in time taken to perform the image manipulation task and memorability of the gestures, respectively.

Usability Study with Surgeons: Last, we conducted another user study with surgeons to compare three interaction modalities: 1. keyboard and mouse; 2. gestures; and 3. speech. Overall, there were 12 surgeons participating in the study, and this study was organized and approved by the Indiana University School of Medicine. Each participant was asked to perform an image manipulation task using all three modalities. The tasks and gesture lexicons (two best lexicons, L6 and L8) and modalities are completely randomized. Results shown in Figure 14 suggests that surgeons consider gesture and speech to be natural and

intuitive; however, there is a legacy bias and a learning curve associated with these modern interfaces. This requires customization and adaptability for these interfaces to be smoothly integrated with the OR.

4. Publications

1. Chanci, D., Madapana, N., Gonzalez, G., & Wachs, J. (2020, December). Correlation Between Gestures' Qualitative Properties and Usability metrics. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 64, No. 1, pp. 726-730). Sage CA: Los Angeles, CA: SAGE Publications.
2. Madapana, N., & Wachs, J. (2020, November). Feature Selection for Zero-Shot Gesture Recognition. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 683-687). IEEE.
3. Madapana, N. (2020, October). Zero-Shot Learning for Gesture Recognition. In Proceedings of the 2020 International Conference on Multimodal Interaction (pp. 754-757).
4. Madapana, N., Gonzalez, G., & Wachs, J. (2020, November). Gesture Agreement Assessment Using Description Vectors. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 40-44). IEEE.
5. Madapana, N., Gonzalez, G., Zhang, L., Rodgers, R., & Wachs, J. (2020). Agreement Study Using Gesture Description Analysis. IEEE Transactions on Human-Machine Systems, 50(5), 434-443.
6. Madapana, N., Gonzalez, G., Taneja, R., Rodgers, R., Zhang, L., & Wachs, J. (2019). Preference elicitation: Obtaining gestural guidelines for PACS in neurosurgery. International journal of medical informatics, 130, 103934.
7. Madapana, N., & Wachs, J. (2019, May). Database of gesture attributes: Zero-shot learning for gesture recognition. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)(pp. 1-8). IEEE.
8. Madapana, N., & Wachs, J. P. (2018, August). Hard zero-shot learning for gesture recognition. In 2018 24th international conference on pattern recognition (ICPR) (pp. 3574-3579). IEEE.
9. Gonzalez, G., Madapana, N., Taneja, R., Zhang, L., Rodgers, R., & Wachs, J. P. (2018, September). Looking beyond the gesture: Vocabulary acceptability criteria for gesture elicitation studies. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 62, No. 1, pp. 997-1001). Sage CA: Los Angeles, CA: SAGE Publications.
10. Madapana, N., Gonzalez, G., Rodgers, R., Zhang, L., & Wachs, J. P. (2018). Gestures for Picture Archiving and Communication Systems (PACS) operation in the operating room: Is there any standard? PloS One, 13(6), e0198092.
11. Madapana, N., & Wachs, J. (2017, November). Zsgl: zero-shot gestural learning. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 331-335).
12. Madapana, N., & Wachs, J. P. (2017, May). A semantical & analytical approach for zero-shot gesture learning. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 796-801). IEEE.

References

1. Bhuiyan, Moniruzzaman, Rich Picking, and others. 2011. Gesture-Controlled User Interface for Inclusive Design and Evaluative Study of Its Usability. Journal of Software Engineering and Applications 4(09): 513.
2. Ebert, Lars C., Gary Hatch, Garyfalia Ampanozi, Michael J. Thali, and Steffen Ross. 2012. You Can't Touch This: Touch-Free Navigation through Radiological Images. Surgical Innovation 19(3): 301-307.
3. Farhadi-Niaki, Farzin, Jesse Gerroir, Ali Arya, et al. 2013. Usability Study of Static/Dynamic Gestures and Haptic Input as Interfaces to 3D Games. *In* ACHI 2013, The Sixth International Conference on Advances in Computer-Human Interactions pp. 315-323. Citeseer.
4. Fukumoto, Masaaki, Yasuhito Suenaga, and Kenji Mase. 1994. "Finger-Pointer": Pointing Interface by Image Processing. Computers & Graphics 18(5): 633-642.

4. Glowinski, Donald, Nele Dael, Antonio Camurri, et al. 2011. Toward a Minimal Representation of Affective Gestures. IEEE Transactions on Affective Computing 2(2): 106-118.
5. Gonzalez, Glebys, Naveen Madapana, Rahul Taneja, et al. 2018. Looking Beyond the Gesture: Vocabulary Acceptability Criteria for Gesture Elicitation Studies. *In* Proceedings of the Human Factors and Ergonomics Society Annual Meeting pp. 997-1001. SAGE Publications Sage CA: Los Angeles, CA
6. Grange, Sébastien, Terrence Fong, and Charles Baur. 2004. M/ORIS: A Medical/Operating Room Interaction System. *In* Proceedings of the 6th International Conference on Multimodal Interfaces pp. 159-166. ACM. http://dl.acm.org/citation.cfm?id=1027962, accessed March 6, 2017.
7. Hurstel, Alexandre, and Dominique Bechmann. 2019. Approach for Intuitive and Touchless Interaction in the Operating Room. Multidisciplinary Scientific Journal 2(1): 50-64.
8. Johnson, Rose, Kenton O'Hara, Abigail Sellen, Claire Cousins, and Antonio Criminisi . 2011. Exploring the Potential for Touchless Interaction in Image-Guided Interventional Radiology. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems pp. 3323-3332.
9. Jost, C., P. De Loor, L. Nédélec, E. Bevacqua, and I. Stanković. 2015. Real-Time Gesture Recognition Based on Motion Quality Analysis. *In* 2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN) pp. 47-56.
10. Kray, Christian, Daniel Nesbitt, John Dawson, and Michael Rohs. 2010 . User-Defined Gestures for Connecting Mobile Phones, Public Displays, and Tabletops. *In* Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services pp. 239-248. MobileHCI '10. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1851600.1851640.
11. Lee, B., P. Isenberg, N. H. Riche, and S. Carpendale. 2012. Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions. IEEE Transactions on Visualization and Computer Graphics 18(12): 2689-2698.
12. Luthra, Vikas, and Sanjay Ghosh 2015 Understanding, Evaluating and Analyzing Touch Screen Gestures for Visually Impaired Users in Mobile Environment. *In* Universal Access in Human-Computer Interaction. Access to Interaction Pp. 25–36. Lecture Notes in Computer Science. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-20681-3_3, accessed July 12, 2017.
13. Madapana, Naveen, Glebys Gonzalez, Richard Rodgers, Lingsong Zhang, and Juan P. Wachs. 2018. Gestures for Picture Archiving and Communication Systems (PACS) Operation in the Operating Room: Is There Any Standard? PLOS One 13(6): 1-13.
14. Madapana, Naveen, Glebys Gonzalez, Rahul Taneja, et al. 2019. Preference Elicitation: Obtaining Gestural Guidelines for PACS in Neurosurgery. International Journal of Medical Informatics 130: 103934.
15. Massaroni, Carlo, Francesco Giurazza, Marco Tesei, et al. 2018. A Touchless System for Image Visualization during Surgery: Preliminary Experience in Clinical Settings. *In* 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp. 5794-5797. IEEE.
16. Mauney, Dan, Jonathan Howarth, Andrew Wirtanen, and Miranda Capra. 2010. Cultural Similarities and Differences in User-Defined Gestures for Touchscreen User Interfaces. *In* CHI '10 Extended Abstracts on Human Factors in Computing Systems pp. 4015-4020. CHI EA '10. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1753846.1754095.
17. Mentis, Helena M., Kenton O'Hara, Gerardo Gonzalez, et al. 2015. Voice or Gesture in the Operating Room. pp. 773-780. ACM Press. http://dl.acm.org/citation.cfm?doid=2702613.2702963, accessed September 13, 2017.
18. Mewes, André, Patrick Saalfeld, Oleksandr Riabikin, Martin Skalej, and Christian Hansen. 2016. A Gesture-Controlled Projection Display for CT-Guided Interventions. International Journal of Computer Assisted Radiology and Surgery 11(1): 157-164.
19. Muller, Michael J., and Sarah Kuhn. 1993. Participatory Design. Commun. ACM 36(6): 24-28.
20. Nacenta, Miguel A., Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of Pre-Designed and User-Defined Gesture Sets. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems pp. 1099-1108. CHI '13. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2470654.2466142, accessed April 20, 2017.
21. Norman, Donald A. 2010. Natural User Interfaces Are Not Natural. Interactions 17(3): 6-10.
22. O'Hara, Kenton, Gerardo Gonzalez, Abigail Sellen, et al. 2014. Touchless Interaction in Surgery. Communications of the ACM 57(1): 70-77.

23. Opromolla, Antonio, Valentina Volpi, Andrea Ingrosso, et al. 2015. A Usability Study of a Gesture Recognition System Applied during the Surgical Procedures. *In* International Conference of Design, User Experience, and Usability pp. 682-692. Springer.
24. Pavllo, Dario, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. *In* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 7753-7762.
25. Piumsomboon, Thammathip, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-Defined Gestures for Augmented Reality. *In* CHI '13 Extended Abstracts on Human Factors in Computing Systems pp. 955-960. CHI EA '13. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2468356.2468527.
26. Rosa, Guillermo M, and María L Elizondo. 2014. Use of a Gesture User Interface as a Touchless Image Navigation System in Dental Surgery: Case Series Report. Imaging Science in Dentistry 44(2): 155-160.
27. Sánchez-Margallo, Francisco M, Juan A Sánchez-Margallo, José L Moyano-Cuevas, Eva María Pérez, and Juan Maestre. 2017. Use of Natural User Interfaces for Image Navigation during Laparoscopic Surgery: Initial Experience. Minimally Invasive Therapy & Allied Technologies 26(5): 253-261.
28. Soutschek, Stefan, Jochen Penne, Joachim Hornegger, and Johannes Kornhuber. 2008. 3-D Gesture-Based Scene Navigation in Medical Imaging Applications Using Time-of-Flight Cameras. *In* 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops pp. 1-6. IEEE.
29. Spagnolo, AM, G Ottria, Daniela Amicizia, Fernanda Perdelli, and Maria Luisa Cristina. 2013. Operating Theatre Quality and Prevention of Surgical Site Infections. Journal of Preventive Medicine and Hygiene 54(3): 131.
30. Spinuzzi, Clay. 2005. The Methodology of Participatory Design. Technical Communication 52(2): 163-174.
31. Stern, H. I., J. P. Wachs, and Y. Edan. 2008. Optimal Consensus Intuitive Hand Gesture Vocabulary Design. *In* 2008 IEEE International Conference on Semantic Computing pp. 96-103.
32. Strickland, Matt, Jamie Tremaine, Greg Brigley, and Calvin Law. 2013. Using a Depth-Sensing Infrared Camera System to Access and Manipulate Medical Imaging from within the Sterile Operating Field. Canadian Journal of Surgery 56(3): E1-E6.
33. Stuij, Sebastiaan Michael .2013. Usability Evaluation of the Kinect in Aiding Surgeon Computer Interaction. PhD Thesis, Faculty of Science and Engineering.
34. Tomasello, Rosario, Cora Kim, Felix R. Dreyer, Luigi Grisoni, and Friedemann Pulvermüller. 2019. Neurophysiological Evidence for Rapid Processing of Verbal and Gestural Information in Understanding Communicative Actions. Scientific Reports 9(1): 1-17.
35. Tsai, Tsai-Hsuan, Kevin C. Tseng, and Yung-Sheng Chang. 2017. Testing the Usability of Smartphone Surface Gestures on Different Sizes of Smartphones by Different Age Groups of Users. Computers in Human Behavior 75: 103-116.
36. Vatavu, Radu-Daniel. 2012. User-Defined Gestures for Free-Hand TV Control. *In* Proceedings of the 10th European Conference on Interactive TV and Video pp. 45-48.
37. Vatavu, Radu-Daniel, and Jacob O. Wobbrock. 2015. 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 pp. 1325-1334. CHI '15. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2702123.2702223.
38. 2016. Between-Subjects Elicitation Studies: Formalization and Tool Support. *In* Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems pp. 3390-3402. CHI '16. New York, NY, USA: ACM. http://doi.acm.org/10.1145/2858036.2858228.
39. Wang, Lily L, James L. Leach, John C. Breneman, Christopher M. McPherson, and Mary F. Gaskill-Shipley. 2014. Critical Role of Imaging in the Neurosurgical and Radiotherapeutic Management of Brain Tumors. Radiographics 34(3): 702-721.
40. Wipfli, Rolf, Victor Dubois-Ferrière, Sylvain Budry, Pierre Hoffmeyer, and Christian Lovis. 2016. Gesture-Controlled Image Management for Operating Room: A Randomized Crossover Study to Compare Interaction Using Gestures, Mouse, and Third Person Relaying. PLOS One 11(4): e0153596.

41. Wobbrock, Jacob O., Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems pp. 1083-1092. CHI '09. New York, NY, USA: ACM. http://doi.acm.org/10.1145/1518701.1518866.
42. Yen, Po-Yin, and Suzanne Bakken. 2012. Review of Health Information Technology Usability Study Methodologies. Journal of the American Medical Informatics Association 19(3): 413-422.