Combining Body Pose, Gaze, and Gesture to Determine Intention to Interact in Vision-Based Interfaces

Julia Schwarz[1], Charles Marais[2], Tommer Leyvand[2], Scott E. Hudson[1], Jennifer Mankoff[1]
Carnegie Mellon University[1], Microsoft[2]
{julia.schwarz,scott.hudson,jmankoff}@cs.cmu.edu, {cmarais,tommerl}@microsoft.com

ABSTRACT
Vision-based interfaces, such as those made popular by the Microsoft Kinect, suffer from the Midas Touch problem: every user motion can be interpreted as an interaction. In response, we developed an algorithm that combines facial features, body pose and motion to approximate a user's *intention to interact* with the system. We show how this can be used to determine when to pay attention to a user's actions and when to ignore them. To demonstrate the value of our approach, we present results from a 30-person lab study conducted to compare four engagement algorithms in single and multi-user scenarios. We found that combining intention to interact with a "raise an open hand in front of you" gesture yielded the best results. The latter approach offers a 12% improvement in accuracy and a 20% reduction in time to engage over a baseline "wave to engage" gesture currently used on the Xbox 360.

Author Keywords: Free-space interaction; vision-based input; user engagement; input segmentation; learned models

ACM Classification Keywords: H5.2 [Information interfaces and presentation]: User Interfaces - Graphical user interfaces.

INTRODUCTION
Recent commercialization of skeletal tracking using depth-sensing cameras holds the promise of bringing free-space, gestural interaction into our everyday lives. One common property that these free-space interfaces share is that they all leverage computer vision to interpret a user's actions (i.e., vision-based interfaces, or 'VIs'). One challenge these VIs present is the Midas Touch Problem: VIs are "always on" and therefore everything a user does may be interpreted as an interaction [12].

This paper addresses the question of *user engagement:* determining when a system should pay attention to a user's actions, and when to ignore them. We describe a machine learning-based algorithm which combines facial features, body pose and body motion to approximate a user's intention to interact, and show how this can be used to make VIs more robust.

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

http://dx.doi.org/10.1145/2556288.2556989

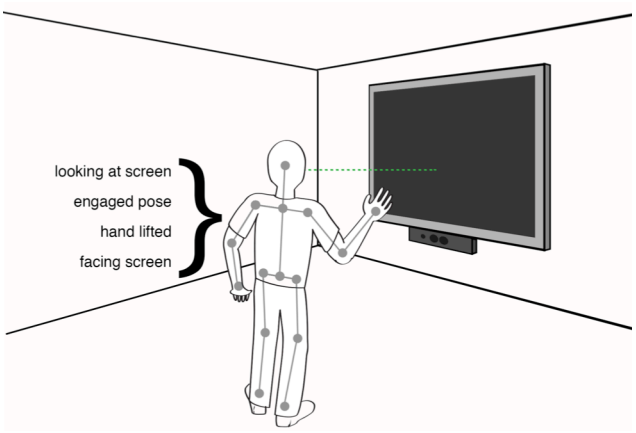

Figure 1. Our system combines results from a collection of individual classifiers to determine a user's intention to interact with a vision-based interface.

To better understand the role of gaze, body pose, and gesture in determining intention to interact, we ran a formative study to identify a set of non-verbal signals people use to communicate intention to interact with vision-based systems, and then trained a set of binary classifiers to recognize these signals. We then used regression analysis to combine these results into an overall intention to interact score, estimating the likelihood of a user's intention to interact (Figure 1).

To illustrate the immediate and practical benefits of using our approach, we built a full-body vision-based interface, and ran a 30-person lab study comparing four different user engagement algorithms in single and multi-user scenarios. We compared an approach that used only our intention to interact score, and two hybrid approaches that combined the intention to interact score with an explicit gesture. Additionally, we include the "wave to engage" algorithm currently used on the Xbox 360 as a benchmark.

Our study results demonstrated that our intention-to-interact metric was useful in determining engagement. Further, by combining this intention-to-interact metric with an easy-to-execute gesture yields even better results, offering a 12% improvement in accuracy and a 20% reduction in time to

engage over the baseline "wave to engage" gesture used on the Xbox 360.

Our contributions are twofold. First, we highlight the value of incorporating information about a user's body pose and motion when determining intention to interact. Second, we show that using an intention-to-interact metric facilitates more accurate and faster engagement detection without requiring users to execute complicated gestures (a common practice today).

RELATED WORK

There has been a wide body of work exploring detection of user engagement, as well as the related question of user attentiveness. These efforts have spanned areas such as desktop interfaces, educational systems, games, driver analysis, human-robot interaction, and human-computer interaction.

For example, Mota estimates a young learner's interest level by analyzing posture data obtained from a pressure-sensitive chair. Mota combined neural networks (posture detection) and Hidden Markov Models to detect engagement at an overall accuracy of 77% [16]. Smith et al. are able to use a single commodity camera to track features such as blinks, gaze direction and mouth movement which can be used to estimate driver's attention level [23]. In the domain of computer interfaces, Asteriadis [1] shows how to perform face detection and tracking to estimate user attentiveness at a computer using only a web camera. Asteriadis is able to correctly detect attentiveness for 88% of a set of 22,000 annotated test frames. Nakano et al. explore conversational attentiveness using gaze transition 3-gram patterns to determine a user's interest in a conversation. They use these insights to develop a multimodal conversational agent which asks probing questions when users are about to lose interest [17]. Our work extends these findings in two ways. First, we address the related question of *determining user engagement*, that is, whether to pay attention to a user's actions, or to ignore them. Second, we demonstrate the value of examining not just the eyes and head but the entire body when determining engagement.

User engagement is important in the domain of human-robot interaction, since a robot must decide whether to pay attention to or ignore a user's actions. Michalowski et al. use a spatial model combined with gaze tracking to determine user engagement with a robot receptionist [14]. Interestingly, the authors observe that in the process of coding the video, they were able to guess a user's engagement by looking at their body motions, suggesting that body motion is a valuable cue to consider. Rich et al. expand on Michalowski's work to include linguistic information such as conversational adjacency pairs and backchannels in their model [20].

In contrast to human robot-interaction, the question of user engagement in single-user desktop interfaces is so trivial that it is almost overlooked. When multiple users are present in a desktop interface, the problem is more complex. One solution is to let each user have control of a finite region [9,24]. When space is limited, several users can share a region [18].

In VIs, a user's body is always tracked. This 'always on' property brings many challenges [12] as users have no natural clutching mechanism [10]. As a result, a person who wants to sip a drink in their living room, for example, may unintentionally trigger music through an inadvertent gesture. This problem is compounded when multiple users are present. When in groups, VIs must know who to pay attention to and who to ignore.

Current VIs use a variety of rules to determine when to begin and end an interaction. For example, the Kinect Dashboard interface on Xbox 360 begins an interaction when it sees a hand wave [15], and ends an interaction when the user leaves the visible area. Several games (see e.g., [8,19]) ask the user to raise one or both hands to begin interacting, and end an interaction either at the end of a game or when the user leaves the scene. StrikeAPose [26] leverages blob tracking to provide a clever "teapot" gesture to detect engagement. Unfortunately, relying solely on explicit gestures is often inaccurate, time-consuming, and not particularly intuitive.

Predictive models have also been applied to detecting engagement. This research largely relies on facial features and gaze tracking to measure interest level. One of the most sophisticated examples is the work of Bohus and Horvitz on predicting intention to interact [4] in open-world dialogue systems [5]. They describe an in-situ model looking at face position and orientation, as well as trajectory. This could be used to predict intention to interact with a kiosk three seconds before an actual interaction with a minimum false positive rate of 5%.

This paper builds off of Bohus et al.'s contribution in two primary ways. First, Bohus' system is primarily intended for spoken dialogue with a machine, whereas our model is intended for non-verbal interactions. The fact that the interactions are all gesture-based, and not voice- or gaze-based, makes the engagement detection problem even more difficult, as there is no input modality change to serve as an additional cue or backup indicator. Second, our intention to interact model incorporates body pose and motion, something Bohus specifically mentions as a potential area of future work.

The use of body pose and motion in determining user's interest level has received additional attention thanks to the advent of inexpensive body tracking systems like Kinect [13]. Bianchi-Berthouse [3] provide an overview of the role of body movement in gameplay, and propose investigating changes in body motion to understand how a user is playing a game. Further, in [2], Bianchi-Berthouse discuss the relationship between body movement and interest level in a game, and show how this relationship can be leveraged to

	# train frames	# test frames	test accuracy	model weight
Engaged stance	189,778	54,359	93.8%	0.67*
Hand lifted above waist	53,182	10,037	98.4%	0.15*
Looking at screen	N/A	N/A	N/A	0.12*
Waving	56,634	16,933	92.5%	0.11*
Hand raised above head	57,344	10,168	92.3%	0.08*
Body facing screen	54,796	7,296	87.4%	-0.05*

Table 1. Test set accuracy of binary classifiers used in our algorithm (train/test data obtained from separate users), as well as weights assigned to each classifier in our final computation. Accuracy refers to per-frame accuracy (# of correct video frames / total # of frames). For example, 98.4% in the hand lifted above waist row indicates that this classifier correctly detected the hand lifted above for 98.4% of gestures. Because "looking at sensor" used built-in face detection algorithm, accuracy numbers are unavailable. Weights marked * are statistically significant (effectively contributed to estimation), $p < 0.05$.

use body motion to enhance gaming experience. In [21], Sanghvi et al. use body lean angle, slouch factor, quantity of motion (computed by looking at pixel-based silhouette motion), and body contraction to determine a student's interest level in playing a game with a robot game companion. Lastly, Kapoor [11] combines facial features, body posture and system state information to determine interest level in a game.

Our work builds on top of the aforementioned research by illustrating how gaze, body pose, and motion can be used in concert in determining a user's intention to interact with vision-based interfaces. Further, we demonstrate the value of combining our intention-to-interact metric with explicit gestures in determining user engagement. This approach offers a new way of thinking about user input: taking account *intention to interact* at the system level. This suggests that future input systems may benefit from directly incorporating uncertainty [22] into the dispatch process.

CHARACTERIZING INTENTION TO INTERACT

To better understand the role of body pose, motion and facial features in determining intention to interact, we ran a formative study exploring how people naturally engage with others, and how they might engage with a vision-based interface. We recruited 16 participants (mean age 25, 2 female, all Caucasian), none of which has experience with VIs.

Participants stood approximately 2.5m away from a VI setup consisting of a depth-sensing camera and 30-inch 1080p display. First, we asked participants to demonstrate how they would get the attention of another person pictured on the display. Second, participants demonstrated how they would "wake up" the system from a sleeping state using gesture, and then pretend to select buttons on a screen (using gesture). Participants repeated each task five times when sitting and standing, and task order was counterbalanced.

We recorded video, depth and skeleton data for each participant in our formative study. One of the authors watched these videos to find common factors that would indicate intention to interact. The following signals were found to be relevant (Table 1): 1) user is looking at the screen, 2) user's hands are lifted above the waist, 3) user is waving one or both hands, 4) user raises a hand above their head, 5) user's body is facing the screen, and 6) user's posture is attentive (arms are not crossed, hands are far away from the head, at least one hand has non-zero velocity).

One feature (one or more hands lifted above waist) was largely due to the type of interaction we were asking users to complete, which was pretending to select buttons in an interface. Not all interfaces have this requirement, indicating that any model that focuses on this feature may be overspecific. This feature, however, can be easily removed in the intention-to-interact model to expand the generalizability of our result to different VIs. The importance of features specific to the interaction at hand should not be understated: in practice interaction-specific features are quite important to generate robust models.

COMPUTING AN INTENTION TO INTERACT METRIC

Based on observations from our formative study, we developed six binary classifiers to detect the body poses and gestures described above.

Our first classifier – *user is looking at camera* – was computed using commodity face detection on the video feed provided from our camera [25]. The remaining five classifiers (*hand is above the waist*, *hand is lifted above head*, *attentive posture*, *body facing sensor*, and *waving*) were trained using hand-labeled frames from a subset of clips gathered during our formative study, as well as clips subsequently gathered by the authors.

We labeled and trained each classifier using Visual Gesture Builder, a tool shipped with Microsoft's Xbox SDK. Visual Gesture Builder uses features such as relative joint angles and joint motion to learn gesture and body pose classifiers using the ADABoost [7] algorithm. Each classifier was trained using ADABoost with 1000 weak classifiers; we also mirrored joint data to account for a lower rate of left-handed users. Body poses were labeled irrespective of whether the user was actually engaged, and we ignored joints below the waist to make classifiers agnostic to whether users were standing or sitting. Table 1 shows statistics regarding accuracy of our binary classifiers. Statistics are presented based on a set of test clips obtained from a

subset of users in our formative study (train and test data obtained from different users).

Our intention-to-interact score is a value between zero and one (inclusive), representing an estimation of the likelihood of a user's intention to interact. The intention-to-interact score is computed using a weighted combination of the six classifiers above; the output of each binary classifier is treated as a zero or one. We used clips recorded from a pilot of our experiment (see below) to get ground truth measurement of engagement for running a regression.

For each clip we recorded the user's ground truth engagement as well as output from our classifiers every 30 msec. In total we collected 16,893 frames over 13 clips. We then ran a linear regression over our features to determine weights that best predicted engagement. Feature weights for the intention-to-interact model are in Table 1. All features significantly contributed to the result estimation ($p < 0.05$), and coefficient of determination (R^2) was 0.66.

We were surprised to see a low weight on hand raised above head and a negative weight on body facing sensor. We posit that this is because the hand raised above head gesture occurred infrequently and that users were almost always facing the sensor regardless of whether they were interacting or not. Interestingly, body pose – not gaze – had the highest weight. This suggests that in the context of VIs (where users may be looking at an interface but have no intention to interact) body pose is more telling than gaze.

In our user study, we used a threshold to determine when a user transitions from *disengaged* to *engaged*, and vice versa. A final component in development of our engagement algorithm was in determining an appropriate threshold: when a user engages and when she disengages.

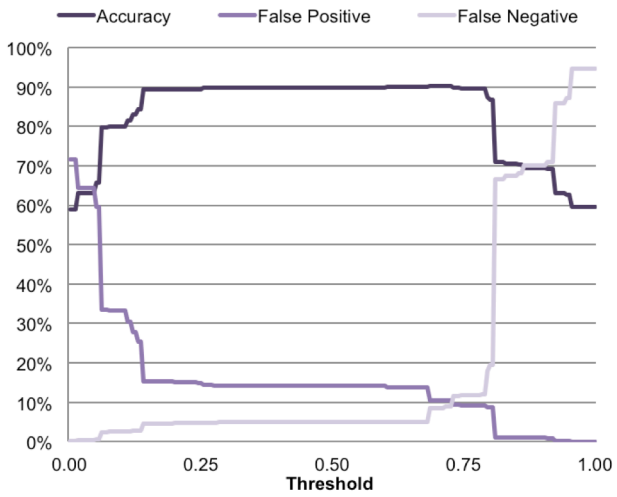

Figure 2. Users transitioned from not engaged to engaged when their intention to interact scores reached a threshold. This figure illustrates the tradeoff between accuracy, false positives, and false negatives at different threshold levels.

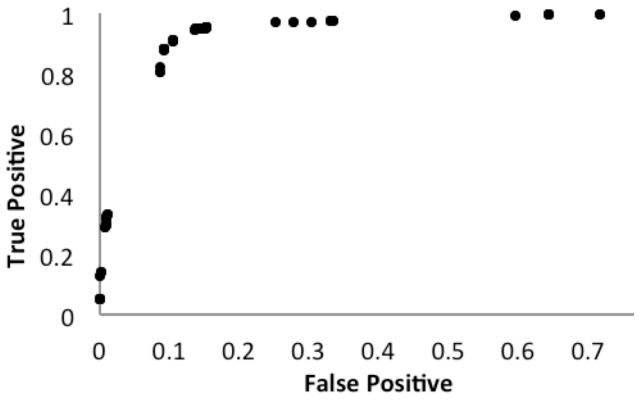

Figure 3. ROC curve for intention to interact model. A' (area under ROC curve) = 0.88.

Figures 2 and 3 illustrate the accuracy tradeoffs of using different thresholds to determine engagement. We used the same 16,893-frame dataset used in computing classifier weights for this analysis. Figure 2 displays curves illustrating tradeoff between false positives, false negatives, and accuracy at different threshold levels, and Figure 3 shows an ROC curve comparing accuracy tradeoff of different thresholds. Based on this analysis, we determined that a threshold between 0.685 and 0.705 offered the highest overall per-frame accuracy in determining intention to interact of 90%, however as will be seen in future sections, this threshold should be adjusted according to the cost of false positive and false negative classifications.

EVALUATION

To demonstrate the value of our intention-to-interact approach, we ran a 30-person lab study comparing three different engagement algorithms, both in single and multi-user scenarios. As a baseline, we used a direct port of the "wave to engage" gesture on the Xbox 360, and compared this engagement method to three of our own methods: a protocol that used only our intention-to-interact score, and two hybrid protocols that combined the intention-to-interact score with a gesture. Our experiment focused on testing the following hypotheses:

H1: *Making use of the intention-to-interact metric leads to higher accuracy when determining user engagement.*

We picked this hypothesis because it tested the overall percentage of time that the system correctly interpreted a user's intention to interact (or not) with a system.

H2: *Making use of the intention-to-interact metric reduces the number of accidental engagements when compared to using only a gesture.*

This hypothesis is useful to test because while overall accuracy is important, a key aspect of any VI (and the core problem behind Midas touch) is reducing accidental engagement.

H3: *Making use of the intention-to-interact metric leads to faster engagement with the system compared to explicit gesture.*

An important consideration regarding the usability of vision-based systems is the ease and speed with which users can begin to use an interface. Therefore, we chose this hypothesis to ensure that the intention-to-interact metric, while accurate, would better enable users to engage with VIs in a timely manner.

H4: *Making use of the intention- to-interact metric leads to faster engagement handoff between users.*

Whether in the living room, kitchen or operating room, VIs often must determine whom amongst a set of possible users to pay attention to. Furthermore, being able to hand off system control between users is also important. We chose this hypothesis to ensure that our intention-to-interact model offered improvements in multi user as well as single user scenarios.

System Description

We built a simple gesture-controlled interface for our study using a pre-release version of the Xbox One Kinect. Software was implemented in C++ and ran on Windows. We used an Xbox One Kinect camera to capture depth images and track skeletons; our system ran at 30 frames per second.

The interface was a simple 'bop the mole' game, where users need to engage and move a hand-controlled cursor over as many targets as possible, then disengage and do some secondary task. Our VI was designed to work for single and multi-user scenarios. Users could also see a small "Picture-in-Picture" (PiP) view of the scene, and get feedback about when they were engaged in the PiP. Users controlled a body-relative screen cursor, one cursor per hand.

Our system recorded depth and skeleton data, as well as results from individual binary classifiers, intention-to-interact scores, and ground truth engagement as determined by the game state. In addition to implementing the classifiers and regression described in the model, we also implemented a direct port of the wave recognizer currently shipped on the Xbox 360 to provide a baseline comparison.

We developed a player engagement protocol to compare algorithms. Our protocol assumes at most one engaged player at any given moment. Engaged users became disengaged when they satisfied a disengagement criteria. If no user was engaged, the player with the highest engagement score (above some threshold) became engaged.

Procedure

We ran two studies to compare our engagement algorithms in different scenarios. The first study was aimed at determining engagement speed and accuracy for a single player. The second study explored a multi-user scenario, specifically, the time to hand off control to another player.

We recruited 30 participants for our single person study (3 female, mean age 32). We used all participants from the single person study in addition to 9 more males and 1 female (mean age 35) for our multi-user study. All participants were professionals in the IT industry and had moderate to high experience with using the Kinect. Participants knew nothing about the internal workings of our algorithm and did not participate in gathering of training data. However, the skewed distribution of participants (mostly male, IT professionals) is a limitation of the study. In particular, our participants may have had prior experience with triggering the "wave to engage" gesture, making it more challenging to demonstrate the benefit of the intention to interact score over the baseline.

The experimental setup was identical for both studies. Participants stood or sat in a living-room sized alcove 2.5m away from a 65" (165.1cm) display, raised 1.5m from the floor. Our Xbox One Kinect camera was located directly underneath the display. Figure 1 illustrates our experimental setup. The experimenter sat in the alcove with participants, giving instructions and moving participants through conditions of the experiment.

Figure 4 provides an overview of the study procedure. In both experiments, participants were asked to repeatedly engage with the system and select a series of buttons, then disengage from the system. To ensure participants tried to engage quickly our game kept a score of the number of buttons selected. In both studies, each participant session was broken into two blocks, one standing block and one sitting block. All conditions were tested in each block. In both studies block and condition order was counterbalanced. For each condition the experimenter explained the engagement method to participants and participants were given one minute to practice.

The experimenter described the wave condition and wave combined with score condition as "wave your hand to engage", the hand lifted combined with score condition as "raise your hand and make sure it is open". During piloting, users appeared to be at a loss for what to do in the score

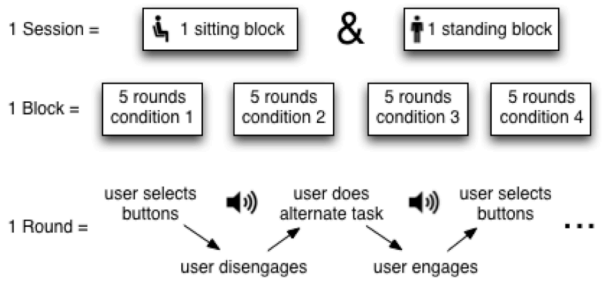

Figure 4. Overview of study design for single user scenario. Each user participated in a single session. Condition and block order were counterbalanced. The multi-person scenario used pairs of participants who were asked to take turns selecting buttons with no alternate task.

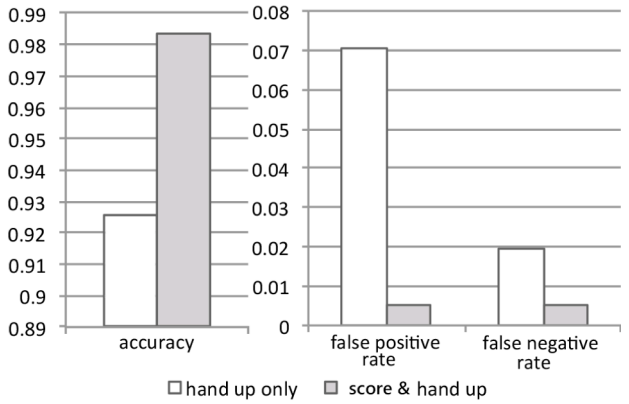

Figure 5. Accuracy and false positive rates comparing a hand raised and hand raised & score algorithm from 3-person pilot study. Error bars omitted due to insufficient data.

only condition unless they were given a specific action to perform, so we told them to "either raise your hand", or "wave".

At the end of each session participants filled out a brief questionnaire asking them which engagement method they preferred. Condition order was counterbalanced to control for order effects.

Study 1: Single Active Player

In our first study, we aimed to compare engagement speed and accuracy across conditions. Within each block, users alternated between selecting buttons on a screen and doing some other task. The latter was randomly selected from a set of tasks designed to mimic what a user may want to do during regular usage. A set of seven alternate tasks were picked from an initial expert brainstorming session, and then these seven tasks were narrowed down to five tasks based on feedback from pilot runs of our study. The tasks were: "Talk on your phone", "Look to the side", "Cross your arms", "Walk over to and move an object from one location to another" and "Browse email on your mobile device."

For each block, users selected buttons for five seconds, and then performed the randomly selected task for five seconds. Users were given audio and visual cues to switch tasks. Each round was repeated five times.

Study 2: Multiple Users

The aim of our second study was to compare time to hand off control to another user between conditions. The high level procedure of the study is described above. Users participated in pairs, giving us a total of 20 pairs. For each pair, each user had a designated color, indicated in the PiP display. In each block users had to engage and select as many of their colored buttons as possible. Button color would alternate every five seconds (indicated by a chime). Users were penalized in the game if they stayed engaged while another person's buttons were visible, thus motivating users to disengage as quickly as possible after selection. This was repeated 20 times per block.

Conditions

For both studies, we compared three engagement algorithms to the baseline wave to engage gesture, resulting in four experimental conditions.

C1: Wave Gesture (control) Our control condition aimed to very closely mimic the "wave to engage" interaction currently shipped on Xbox 360. For this we implemented a direct code port of the wave recognizer on Xbox 360. The engage criterion for this condition was recognition of the Xbox 360 wave gesture, and disengagement occurred 250ms after both of the user's hands were below the waist.

C2: Score only This condition used only our intention-to-interact score to determine engagement and disengagement. The engagement criterion for this condition was based off of a smoothed intention-to-interact score. We smoothed the intention-to-interact score in our implementation using (for simplicity) an exponential smoothing filter [6] with a smoothing factor of 0.8, applied every frame. Our smoothing factor was determined from observations during pilot runs of the study.

Users became engaged when their intention-to-interact score rose above 0.69. This threshold was found to provide a relatively low false positive rate of 10% while maintaining an overall accuracy (on our training dataset) of 95% (see Figure 2). Users became disengaged when their intention-to-interact score dropped below 0.1, which offered a nice tradeoff between false positives, true negatives and false negatives.

C3: Score Combined with Wave Gesture Another way to reduce false positive engagement is to add an additional gesture requirement. The score combined with wave condition used the same threshold of 0.69, this time with an unsmoothed score, and also required that the user complete the Xbox 360 wave gesture. Users disengaged when their unsmoothed intention-to-interact score dropped below 0.1.

C4: Score Combined With Hand Raised and Open We were also curious to see if an easy-to-execute, but error-prone gesture ("hand raised and open") could be combined with the intention to interact score to create a fast yet accurate engagement protocol. In this condition we required the unsmoothed intention-to-interact score to again be above 0.69 and that at least one of the user's hands was above the waist and open.

We used a hand state classification algorithm similar to what is shipped in the Kinect for Windows SDK v1.7 for hand state detection. Disengagement was the same as in C3.

To verify our intuition that the hand raised and open gesture would generate too many false positives to be practical, we ran a three-person pilot of the full study described below on six runs of three different people (total of 1,785 frames), hand labeled these frames according to whether the user was engaged with the system, and compared accuracy and false positive rates (see Figure 5).

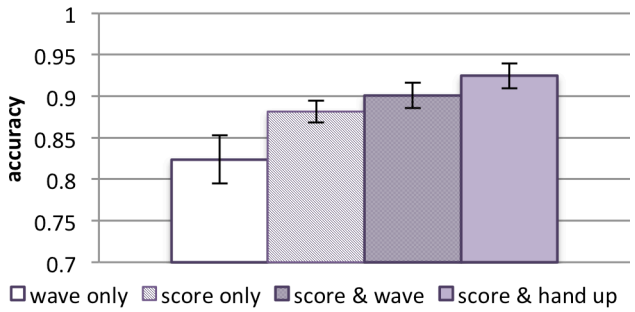

Figure 6. Mean of per-frame accuracy across users, single person study. Bars represent 95% CI.

Using just hand up and open had an accuracy and false positive rate of 93% and 7% respectively, while the combined model had an accuracy and false positive rate of 98% and 0.5%. Results from this preliminary study indicate that this simple gesture, while easy and fast, was not accurate enough for detection of user engagement, and that an additional metric such as our intention-to-interact score would be necessary to better determine user engagement.

Data

During our study, we recorded depth and skeleton data for all runs, and recorded binary classifier results, as well as an engagement score every 300ms. We were able to automatically determine ground truth because the study interface controlled programmatically the user's engagement state (Figure 4). During the study the experimenter made sure that users were on task and switched tasks in a timely manner.

In total, we recorded 1305 clips, representing 191,905 frames of data, 1,200 engagements, 1,200 disengagements, and 3,200 engagement handoffs across the single and multiple person study. Two users from our single person study had corrupt data or failed to complete the study in an attentive manner and were removed from analysis. To ensure that all data samples were independent, we performed all analysis on a per-user basis. The statistics presented below are based on 28 per-user samples in our single person study, and 20 per-pair samples in our multiple person study.

As ground truth we recorded the each user's expected engagement, based on the task that they were currently doing. In the single person study, the ground truth was based on whether users were supposed to be performing button selection or the distractor task (see Figure 4). In the multiple person study, the ground truth was based on which user was asked to select the buttons. To account for user reaction time (the time between when the chime sounds and the user actually starts switching tasks), we ignored all data one second after a chime in all conditions. In the analyses presented below, all data were normally distributed, or log transformed to become normally distributed.

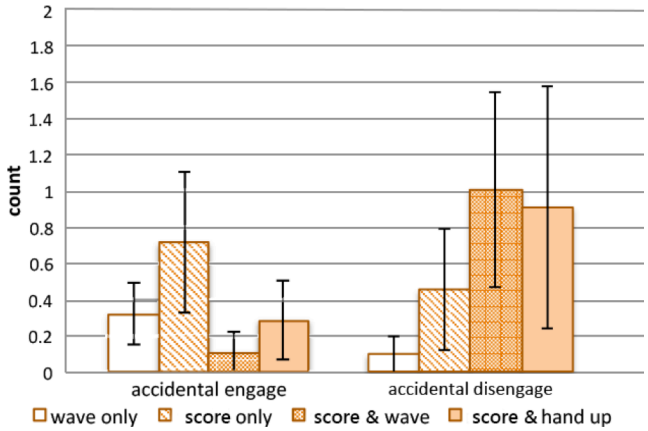

Figure 7. Mean number of accidental engagements and disengagements per user in single person study. Bars represent 95% CI.

Results and Discussion

Our hypotheses were largely confirmed, with the exception of H2, which was partially confirmed. Below we describe these results with respect to each specific hypothesis.

H1: *Making use of the intention-to-interact metric leads to higher accuracy when determining user engagement.*

Our results indicate that using an intention-to-interact score leads to higher accuracy over the baseline wave to engage condition. All analyses for this hypothesis were performed on data from the single person study. To determine impact of condition on overall accuracy, we used several measures.

First, for each user we measured per-frame accuracy (# frames correct / total number frames) after ignoring all frames within one second of a chime. This number was chosen to ensure that the cutoff was considerably less than the fastest possible engagement method. However, this metric unfairly biases situations where users are accidentally engaged for a long time (i.e. they fail to disengage). Therefore, we also counted the number of accidental engagements and accidental disengagements per user for each condition (out of a total of 40 engagements and 40 disengagements per user). Accuracy was computed on a per-user basis. Figure 6 shows results comparing overall accuracy, while Figure 7 show results comparing number of accidental engagements and accidental disengagements in our system.

Using our engagement score combined with a hand up and open gesture yielded the highest per-frame accuracy, at 92.4%, while wave only was at 82.3%. We log transformed our per-frame accuracy to normalize the distribution. A repeated measures ANOVA on the log-transformed data comparing per-frame accuracy indicated a difference across conditions ($F(3, 81)=21.68$, $p < 0.01$), and a post-hoc test applying Bonferonni correction indicated that the three methods leveraging the intention-to-interact score had higher mean accuracy than wave only ($p < 0.01$ in all cases).

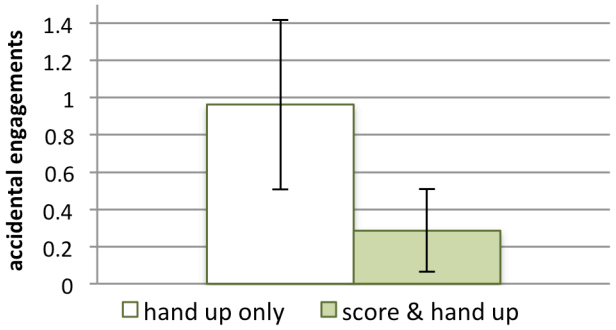

Figure 8. Post-hoc comparison of accidental engagements with and without an intention to interact score (single person study) when using a hand up and open gesture.

There are two possible explanations for this difference. First, the wave gesture takes a considerable time to engage. Also, the disengage condition for our wave only method caused a failure to *disengage* in some cases. This failure to disengage was a surprising result, which caused accuracy to drop significantly for the wave gesture. To reduce the impact on prolonged failures to disengage, we also counted the number of accidental engagements per user in our single person session. Figure 7 compares number of per-user accidental engagements and disengagements. A detailed analysis of accidental engagement will be presented in the next section.

One interesting observation is that our intention-to-interact score led to more accidental disengagements than using an explicit gesture. This is not surprising as our intention-to-interact metric measures *intention to interact*, and not *intention to disengage*. In future work we suggest developing a separate *intention to disengage* metric to reduce accidental disengagement.

H2: *Combining intention-to-interact with an explicit gesture reduces the number of accidental engagements when compared to using only an explicit gesture.*

Our study did not show that the combining of a wave gesture with our intention-to-interact score helped reduce accidental engagement (Figure 7, left). One likely reason for this is that the wave gesture was already specifically tuned to reduce false engagement, at the cost of other factors such as engagement speed (see *H1*). Additionally, the number of accidental engagements per user was already very low in all conditions, at less than 0.8 false engagements out of 40.

While the intention-to-interact score did not perform significantly better than the wave gesture, a post-hoc analysis showed that the engagement metric does reduce false engagement when using the raise hand up gesture.

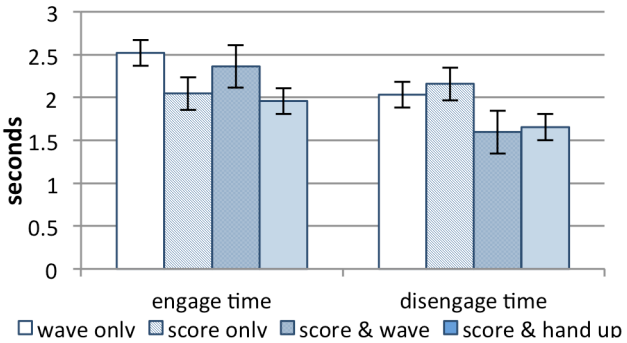

Figure 9. Mean time to engage, disengage in single person study, across users. Bars represent 95% CI.

For our post-hoc analysis, we simulated a "hand up and open" engagement gesture by replaying depth and skeleton data collected from our study, and this time used hand up and open as the engagement criteria, hands dropped as disengagement criteria. We recorded binary classifier results, ground truth, and computed engagement. One recording from our study was thrown out due to corrupt data, yielding a total of 27 data points per condition.

Figure 8 shows results of our post-hoc comparison. A *t*-test indicated that the intention-to-interact score did lead to significantly fewer accidental engagements between the hand up only condition and the intention-to-interact score combined with hand up condition ($t(26)=2.62$, $p < 0.05$). Specifically, accidental engagement dropped from 0.96 accidental engagements in the open hand up only condition to 0.29, a 70% reduction in accidental engagements.

H3: *Making use of the intention-to-interact metric leads to faster engagement with the system compared to explicit gesture.*

Analysis of data from our single person study indicates that using an intention-to-interact score leads to faster engagement time than the baseline wave to engage gesture. Specifically, average time to engage with wave was about 2.5 seconds while average time to engage in the score-only condition and the score combined with hand up condition was about 2 seconds, yielding a 20% reduction in engagement time (Figure 9).

A repeated-measures ANOVA across all conditions indicated a statistically significant difference between average engagement times ($F(3, 81) = 9.66$, $p < 0.01$). Post-hoc tests applying Bonferroni correction indicated a statistically significant difference in time to engage between using the wave gesture and our score only at $p < 0.05$, as well as a difference between wave only and hand up & score ($p < 0.05$). The fastest engagement method was obtained using intention-to-interact combined with a hand up and open.

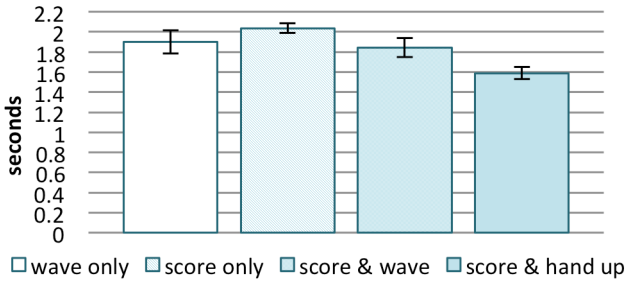

Figure 10. Mean time to hand off control in multi-person study, across users. Error bars represent 95% CI.

H4: *Using an intention-to-interact score leads to faster engagement handoff between players.*

For this analysis, we only used data from our multi-person study. Surprisingly, handoff time was slowest in the intention-to-interact score only condition (Figure 10). We posit that this was because time to disengage was quite high in the intention-to-interact score only condition (due to smoothing of the intention-to-interact score). Since our scenario required that the currently engaged user first disengage before another user can take control, a high disengagement time would increase handoff time as well. The high disengagement time in our intention-to-interact only condition was a result of aggressive smoothing of the intention-to-interact score.

Nevertheless, this study once again illustrated the benefits of combining the intention-to-interact score with a hand up and open gesture. Specifically, we found that handoff time in the intention-to-interact score combined with hand up condition was significantly lower than handoff time in the wave only condition. Time to switch dropped from 2.04 seconds (wave) to 1.59 seconds, a 22% reduction. A repeated measures ANOVA confirmed a significant difference across conditions ($F(3, 57) = 28.40$, $p < 0.01$), and post-hoc tests using Bonferroni correction confirmed a difference between wave only and hand up combined with score at $p < 0.05$. Note that the degrees of freedom in this analysis are different from previous analyses because we are analyzing results from the multiple person study, which had 4 conditions, and 20 pairs of users completing each condition.

Subjective Gesture Preference

Questionnaire data collected during both studies showed a strong preference for the hand up and open gesture compared to the wave gesture (Figure 11). Interestingly, people strongly preferred the hand up combined with intention-to-interact score condition to the score only condition. One explanation for this could be that in the score only condition users did not know exactly what they should do and thus could not be sure that their actions would lead to an engagement. This further speaks to the need of always combining an intention-to-interact score with some gesture so that users can always know what action will trigger engagement.

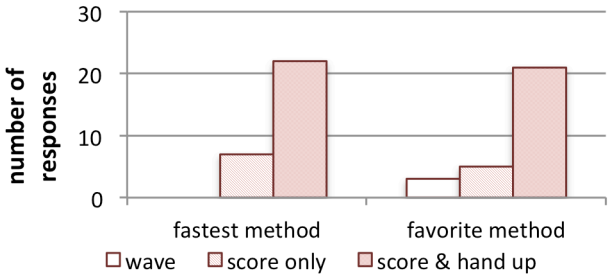

Figure 11. Gesture preference from questionnaire (single person study).

Limitations

Our studies and subsequent analysis demonstrate the benefit of using an intention-to-interact metric for determining intention to interact with vision-based systems. Nevertheless, there are a few limitations we should note.

Foremost, we smoothed our model score considerably to reduce false engagements and disengagements, which caused an unusually high disengagement time. However, even with aggressive smoothing we were able to show benefit over our baseline. Secondly, our baseline engagement, to which we compared, may not be the fastest or most accurate method available across all of the research literature. We chose it primarily because it was an example of a commercially available, in-the-wild engagement detection algorithm that was highly tuned and tested.

CONCLUSIONS AND FUTURE WORK

In this paper, we argue for the inclusion of an intention-to-interact metric when determining user engagement in vision-based input systems. We used a collection of binary classifiers that looked at a user's body pose, motion, and gaze to approximate the likelihood that a user intends to interact with an interface.

Results from two 30-person lab studies indicate that while an intention-to-interact metric by itself is useful, the full power of our intention-to-interact score becomes evident when combined with an easy-to-execute but error-prone activation gesture. Our intention-to-interact metric allows vision-based input systems to accurately detect user engagement without requiring overly complex gestures.

This work opens up several avenues of future work. First, our model for intention to interact focuses largely on detecting engagement with the system. However, the model could be further improved if we included intention to disengage. Second, we have demonstrated the value of using body pose and gesture in determining intention to interact, and believe that these findings are relevant to other types of vision based scenarios, such Human-Robot Interaction and dialogue management systems. Finally, the model offers a new way of thinking about and segmenting user input: generating an *intention-to-interact score* and dispatching user input according to this score, suggesting that future input systems

may benefit from directly incorporating uncertainty [22] into the dispatch process.

ACKNOWLEDGMENTS

The authors thank Chris Harrison for his assistance, as well as the members of the Xbox NUI Team for their feedback and support. This work was funded by a Microsoft Graduate Research Fellowship and NSF Grant IIS1217929. Icons from Figure 4 were by Jessica Lock and Mellonie Manohar from The Noun Project.

REFERENCES

1. Asteriadis, S., Karpouzis, K., and Kollias, S. Feature extraction and selection for inferring user engagement in an hci environment. *Human-Computer Interaction*, Springer-Verlag (2009), 22–29.
2. Bianchi-Berthouze, N. What Can Body Movement Tell Us About Players' Engagement? *Measuring Behavior '12*, ACM Press (2012), 94–97.
3. Bianchi-Berthouze, N. Understanding the role of body movement in player engagement. *Human Computer Interaction 28*, 2 (2013), 42–75.
4. Bohus, D. and Horvitz, E. Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. *SIGDIAL '09*, ACM Press (2009), 244–252.
5. Bohus, D. and Horvitz, E. Dialog in the open world: platform and applications. *ICMI-MLMI '09*, ACM Press (2009), 31–38.
6. Brown, R.G. *Exponential Smoothing for Predicting Demand*. Little, 1956.
7. Freund, Y. and Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences 55*, 1 (1997), 119–139.
8. Harmonix. Dance Central. 2012. http://www.dancecentral.com.
9. Hartmann, B., Morris, M.R., Benko, H., and Wilson, A.D. Pictionaire: supporting collaborative design work by integrating physical and digital artifacts. *CSCW '10*, ACM Press (2010), 421–424.
10. Hinckley, K., Pausch, R., Goble, J.C., and Kassell, N.F. A survey of design issues in spatial input. *UIST '94*, ACM Press (1994), 213–222.
11. Kapoor, A., Picard, R.W., and Ivanov, Y. Probabilistic combination of multiple modalities to detect interest. *ICPR '04.*, IEEE Computer Society Press (2004), 969–972.
12. Kjeldsen, R. and Hartman, J. Design issues for vision-based computer interaction systems. *PUI '01*, ACM Press (2001), 1–8.
13. Kleinsmith, A. and Bianchi-Berthouze, N. Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing 9*, Preprint (2012).
14. Michalowski, M.P., Sabanovic, S., and Simmons, R. A spatial model of engagement for a social robot. *Advanced Motion Control '06*, IEEE Computer Society Press (2006), 762–767.
15. Microsoft. Xbox 360 + Kinect. http://www.xbox.com/kinect.
16. Mota, S. and Picard, R. Automated posture analysis for detecting learner's interest level. *In CVPR Workshop on HCI*, (2003).
17. Nakano, Y. and Ishii, R. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. *IUI '10*, ACM Press (2010), 139–148.
18. Pawar, U.S., Pal, J., Gupta, R., and Toyama, K. Multiple mice for retention tasks in disadvantaged schools. *CHI '07*, ACM Press (2007), 1581–1590.
19. Rare. Kinect Sports. 2010. http://www.rareusa.com/games/kinect-sports.
20. Rich, C., Ponsler, B., Holroyd, A., and Sidner, C.L. Recognizing engagement in human-robot interaction. *HRI '10*, IEEE Computer Society Press (2010), 375–382.
21. Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P.W., and Paiva, A. Automatic analysis of affective postures and body motion to detect engagement with a game companion. *HRI '11*, ACM Press (2011), 305–312.
22. Schwarz, J., Hudson, S.E., Mankoff, J., and Wilson, A.D. A Framework for Robust and Flexible Handling of Inputs with Uncertainty. *Proceedings of the 24th annual ACM symposium on User interface software and technology UIST '11*, ACM Press (2010), 47–56.
23. Smith, P., Shah, M., and da Vitoria Lobo, N. Determining Driver Visual Attention With One Camera. *IEEE Trans. on Intelligent Transportation Systems 4*, 4 (2003), 205–218.
24. Stødle, D., Hagen, T., Bjorndalen, J., and Anshus, O. Gesture-based, touch-free multi-user gaming on wall-sized, high-resolution tiled displays. *Journal of Virtual Reality and Broadcasting 5*, 10 (2008), 1860–2037.
25. Viola, P. and Jones, M.J. Robust Real-Time Face Detection. *International Journal of Computer Vision 57*, 2 (2004), 137–154.
26. Walter, R., Bailly, G., and Laboratories, T.I. StrikeAPose : Revealing Mid-Air Gestures on Public Displays. (2013), 841–850.