# Bartertown: a single-player human computation game to create a dataset of iconic gestures

Wouter van den Heuvel
Graduation Thesis, June 2015
Media Technology MSc program, Leiden University
Supervisors: Maarten H. Lamers and Amir Sadeghipour (Medical University of Vienna)

**Abstract —** This work explores the use of the *human computation games* paradigm to create a dataset of iconic gestures describing primitive 3D shapes. To this end we have created *Bartertown*, a single-player video game using Microsoft Kinect v2. Contrary to many human computation games, our game is not just about providing annotations to existing data, instead players are the primary data providers. Data quality is achieved by virtue of a self validating system. We assert that the medium of games can be employed successfully in these scenarios. We reflect on the prototype and propose a list of general guidelines useful for researchers interested in creating human computation games for acquisition of gesture corpora.

Keywords: gestures, gesture recognition, human computation, games with a purpose

## I. INTRODUCTION

Research and development in virtual environments and new, less cumbersome modes of human computer interaction (HCI) towards the end of the last century have sparked interest into modelling, analysing and recognition of human gestures [1]. Robot control [2] and immersive gaming [3] are other possible applications that could benefit from gesture recognition. Finally, one could imagine computers understanding human gestures could be useful in the reversed scenario; that is, virtual agents or even robots expressing meaning and affect through gestures, thus increasing the acceptance of social and affective characteristics [4], [5]. It is well known that gestures have a positive effect on teaching [6], [7] and will simultaneously improve student's perception of teachers [8]. This means that gestures can be useful in a computer aided learning scenario. Systems that use gestures could be developed for assisting the hearing impaired. Research has discovered that gestures without speech can assume the full burden of communication and take on a language-like form [9]. This could be very useful in overcoming the language barrier in robot-human communication with foreigners or young children who have yet to learn a (spoken) language.

Computer scientists have turned to different statistical models for automated gesture recognition, e.g. Bayesian Networks [10], Hidden Markov Models and Finite State Machines [11]. For all these models to work the algorithm needs to be trained with a training set, i.e. a corpus of data [10], [12]–[14]. Using these training sets, an algorithm can learn to recognise gestures performed by humans. Because the set of iconic gestures is so large and the differences in gesture performance between individuals can vary greatly [15], the training set has to have sufficient size for any gesture recognition system to work, both in number of gesture classes and variations therein.

Traditionally, researchers have relied on gathering this data themselves, resorted to volunteers (often people in their immediate, academical surroundings) or outsourcing the task externally. Other than having a small, often skewed demographic, this method of data acquisition is not very efficient in terms of resources and manpower. Particularly because of the variation in gesture execution research requires many different gesticulators each performing a great number of gestures. For the research described in this paper, we turn to a different solution.

Gestures can be performed and recognised with relative ease by the majority of the population. This makes performing gestures and providing labels (ground truth) to a collection of video images an ideal candidate for crowdsourcing [16]. The problem then turns into a motivational one, '*how to get the crowd motivated to input data?*' An existing method is using a micro-task market service where workers get financial compensation, e.g. Amazon Mechanical Turk [17]. Some research was done into crowdsourced annotation [18]. Because of the inherent noisy nature of crowdsourcing, it is imperative to have a large volume of annotators, thus driving up research costs.

Thanks to von Ahn's pioneering research [19] there exists another approach, so called human computation games, or: 'games with a purpose'. The principle is simple, one creates a computer game that is fun to play for people and at the same time the players (sometimes even unwittingly) provide useful data for tasks that computers cannot yet perform. Aside from having lower costs, we also believe that because workers will be motivated by having fun rather than financial gain this process will lead to higher data quality. This is, however, still an assumption and no substantial research into this hypothesis was found. The games are set up such that correct 'calculations' lead to a higher player score. Arguably the first of these games was the *ESP game* [19] which labels images, but many researchers have applied this paradigm to different problems (see II.c On human computation and games with a purpose).

Although every instance of a human computation game is different, two aspects are always present: enjoyability and data quality control. More people playing means more and better data and for this reason the game has to be widely available and appeal to a large audience enough for them to invest time in it. Human computation games that are enjoyable to play have proved to be very popular indeed [20] and thus successful in that regard. Another important aspect of human computation games is safeguarding the system against people who might cheat or otherwise pollute the data. Nevertheless, a number of countermeasures exist and have been successfully applied [21]–[23].

Considering all of the above, we ask ourselves the question: *Can we develop an entertaining video game for the purpose of creating a dataset of usable iconic gestures?* We expect a database with a large volume of high quality data to be highly useful to scientists in training new and/or testing existing machine learning models. These trained models could be useful in scenarios of affective computing, online education and instructional material, conceptual design [24], home automation, gaming, etc. The dataset on its own could also prove valuable for anyone doing quantitative analysis on how we use and interpret gestures. The reader should take note that any actual recognition algorithms are outside of the scope of this paper.

The remainder of this article is organised as follows. Section II reviews prior research on gestures, gesture modelling and recognition, gesture datasets and finally human computation and games with a purpose. Section III discusses the properties of the gesture dataset. Section IV details the development of the game and test set-up. Section V offers insight on the most important results. Section VI proposes a number of guidelines and design principles for human computation games for collecting gesture corpora. Section VII provides a discussion and conclusion of our work.

# II. RELATED WORK

## II.a On gestures in communication and their semantic aspects

The gestures we make during speech (often involving the arms and hands) have been an area of research for decades. Currently, there are a number of gesture classification schemes in use. Ekman and Friesen [25] and Efron [26] have suggested different taxonomies. McNeil [27] defined a set of high level categories based on the referential characteristics of gestures which is useful in our case. One of these categories is *iconic,* "hand gestures that represent meaning that is closely related to the semantic content of the segments of speech that they accompany" [28]. In contrast to *emblematic gestures*, that have a conventionalised and often culture-specific form and meaning (e.g. the 'thumbs up' symbol), iconic gestures are performed spontaneously and have no codified semantics attached. Notwithstanding, they are widely understood by humans across many cultures and their number is virtually inexhaustible. For this reason we have chosen to focus on this category of gestures. From this point on, when we refer to 'gestures' in this article we mean iconic gestures.

Gestures have ostensibly been considered to communicate, in parts, the message of the accompanying speech [29]. Hadar *et al* [30] have researched into this concept of *semantic specificity*, "the clarity or non-ambiguity with which a particular gesture indicates the meaning associated with it". Hadar *et al* found that humans are able to select the right meaning of a gesture in a multiple-choice scenario [30], but when trying to determine the intended meaning of a gesture on its own humans perform barely better than chance [30], [31]. Because we will not be giving any conversational context in our game scenario it is important to regard the limitations on human capabilities to recognise another persons' gesturing.

## II.b On data sets for human gesture recognition

Ruffieux and Lalanne have recently reviewed a list of currently available datasets for human gesture recognition [32]. They considered aspects like what type of sensor and what view (i.e. front view, top view), how many subjects and if they were standing or sitting, how many 'gesture classes' (i.e. the semantic descriptor of a gesture) and 'instances' (i.e. the gesture performed by a person) and video resolution. A number of guidelines were determined for creating a useful human gesture dataset:

(1) Careful design—before implementing all features and recording conditions should be defined.
(2) Software development—there exist a number of frameworks to record datasets, for more complex scenarios it might be required to write custom software.
(3) Acquisition methodology—should define the process of the acquisition and labelling of the data and ground truthing.
(4) Acquisition—requires rigorous testing in real conditions beforehand. Video should be captured in the highest possible data quality and then optionally compressed for distribution.
(5) Annotation and Verification—data should be annotated and verified via algorithms or manually.
(6) Documentation—the entire acquisition set-up and data should be precisely described if the dataset is to be released publicly.

Sadeghipour *et al* have compiled a dataset of iconic gestures referring to physical objects using 29 subjects [33]. Although the subjects had to gesticulate rather simple shapes, the techniques used for gesturing were very different because each subject was free to perform their own gesture to depict each

shape. Two types of variations were observed, inter-class variations (e.g. changes in direction, velocity, degree of simplification) and intra-class variations (e.g. pantomiming biting an apple vs. drawing the contours of an apple). Fothergill *et al* express the quality of a gesture dataset along two dimensions, *correctness* and *coverage [34]*. It is found that when a gesture class is presented to a group of subjects in a textual modality, coverage of the gesture will be high. We suspect that intra-class variations (as defined by Sadeghipour [33]) are an important component of coverage.

### II.c On human computation and games with a purpose

Conceived by von Ahn, the ESP game [20] was a web based game where two players would be randomly paired and shown the same image. The players could not communicate with each other but could type descriptions of the image. If both players typed the same word, that would mean that word is somehow related to the image and so a database of labelled images could be created. This idea was expanded though the game Peekaboom [21], where the labelled image collection could be defined with finer granularity by players connecting semantic meaning to specific areas within the image.

The concept proved fruitful because soon after a large number of human computation games were developed across a wide array of disciplines. Some are semantic tagging games, e.g. *TagATune [35]* that labels an audio database, *GalaxyZoo* for morphological classification of galaxies and *Guess Who*? [36] for affective facial expressions or *Verbosity* [37] to collect common sense knowledge. Some implementations are quite advanced in their design and facilitate scientific discovery, e.g. *FoldIt [38]* that helps computing protein structure, *EyeWire* [39] to map neurons in the brain or even quantum optimisation in the case of *Quantum Moves [40]*. These examples are to give the reader an idea of the broad scope in which the human computation paradigm has been applied to and is by no means an exhaustive list.

A number of categories have been devised to classify human computation games [16], although this taxonomy could be somewhat outdated as some of the newer games do not fall within these descriptions. Nevertheless we will give a brief explanation of the different types of human computation games:

*(1) Output agreement* games have to be played with two players at the same time. They are paired randomly and unable to communicate. They are presented with an input (often an image, video or sound) and have to provide some output. Whenever the output of both players match, both players win. The game now knows that two players independently agreed on an output, so is very well suited for labelling objective data.

*(2) Input agreement* games are also played by two randomly paired players who receive an input (e.g. an audio sample). However the input they receive could be different from each other. The players are able to describe their input to the other player. Both players then have to indicate whether they think both were exposed to the same input. This type of game is useful for collecting subjective data.

*(3) Inversion problem* games have players in asymmetric roles, one player receives an input and has to provide hints to the other player. When the other player correctly guesses the input the output provided by the first player is assumed to be relevant to the input.

# III. DATASET

Our objective is to compile a dataset containing labelled gestures, expressed as lists of 3D vectors representing position and rotation of limbs in space along a time axis. This data could then be directly used as a training set for a machine learning algorithm or replayed by a virtual character model on screen and studied by researchers directly. The data will be recorded using a Microsoft Kinect V2. Gesture acquisition as well as annotation will be done by the players of the game.

## III.a Size and scope

We have compiled a list of 8 different primitive shapes aiming for maximum visual distinctness between them. These are our gesture classes (see figure 1). In daily non-verbal communication, simplifying complex objects and referring to their abstract shapes is a commonly used strategy while performing iconic gestures. [41] For this reason we believe a collection of iconic gestures referring to basic shapes can be useful for researchers and developers in different scenarios. Our target amount of participants is 30, who will all be adding 4 gesture instances, resulting in an average of 15 gesture instances per class. A high instance per class ratio is required for future use in machine learning algorithms [32]

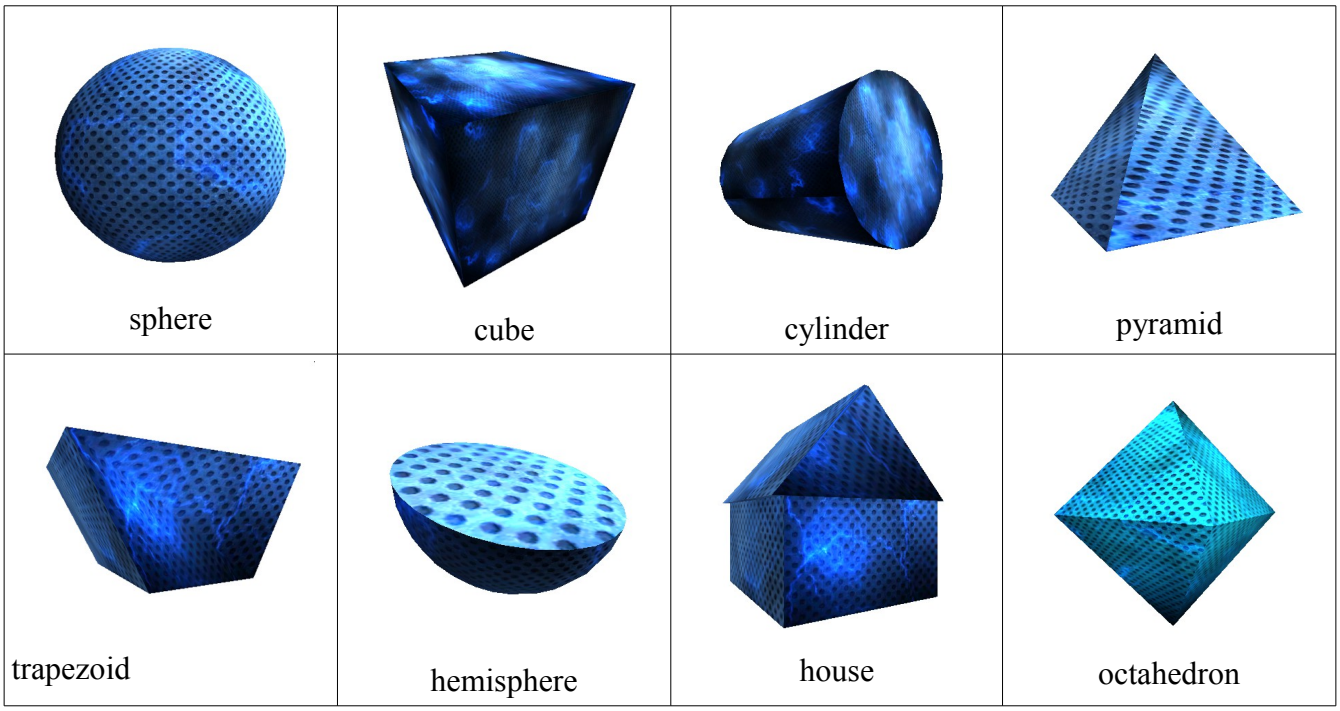


**Figure 1**: Gesture classes, represented as textured 3d primitives.

The dataset contains all gesture data acquired as well as information regarding ground-truth labelling. This means researchers have insight in who labelled which gesture with which label and which alternatives were available. We have also opted to include subject meta-data, e.g. gender, age, nationality and handedness [33]. A full description of all data fields is included in a text file that ships with the dataset. Participants will be asked to perform the gestures standing, facing the camera. The camera will have a full body, unobstructed view of the subject.

## III.b Technical specifications

The Kinect for Windows V2 SDK API can identify up to 25 joints per tracked body. See figure 2 for a diagram. We will record all tracked limbs position and rotation for every frame, at a sampling frequency of 30 frames per second. Machine learning algorithms need to generalise over many different performances so homogeneity is important. For this reason we discard the gesturer's body position in 3D space, only the position and rotation of the joints relative to the root joint are of importance. Because we constructed our gesture vocabulary such that concepts will not have to be expressed through facial expressions or complex finger gestures we expect the data Kinect records will be of sufficient quality.

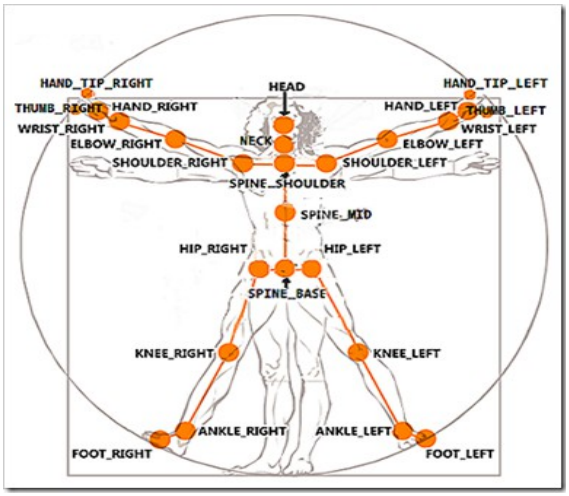


***Figure 2**: Kinect V2 Body (source: Kinect for Windows SDK documentation).*

Our dataset has complexity not present at other gesture datasets, resulting from the fact that both gesture acquisition and ground truth labelling is performed by untrained personnel. This data as such can be interesting when, for example, researching human ability to recognise gestures. Another different aspect of our dataset is that our gesture data ground truth is variable. This results in confidence values that could be interesting to investigate. Because of this complexity we have chosen to store the data in a relational database. This will provide the flexibility to aggregate various aspects of the data. We use a SQLite database because it is public domain software, very portable and easy to integrate. For convenience, we will also provide a flat comma separated file download.

# IV. EXPERIMENT

## IV.a Prototype development

A prototype game, titled “Bartertown” was developed in Unity 4.6[1], a cross-platform game engine very suitable for rapid development of 3D games. The majority of human computation games are multi-player [16] and thus require an initial critical mass of players to be successful [42]. Because we are using specific hardware to record gestures we opted to create a single-player game setting, alleviating the cold start problem of requiring a number of players playing simultaneously. Special care was taken to provide appealing visual aesthetics and an interesting story line to enrich the play experience. A science-fiction setting was chosen, a time-honoured tradition in video games. Research [43] has shown that in human computation games, a plausible story helps create immersion and takes away a subject's feeling of performing labour. For this reason, the prototype included a short intro sequence, as is custom in virtually all video games popular today.


***Figure 3**: Game screen captures of intro sequence (left) and in-game scene (right)*

In the intro, the player sees a spaceship stuck in a meteoroid storm and having to crash land. The game informs the player that in order to repair her ship, she must find eight spare parts, scattered throughout the world. The locations of these parts are known to an alien creature nearby, but in order to communicate with the creature, the player must use gestures (because the creature obviously speaks a different language). Screen captures are shown in Figure 3.Thus, the first spare part is shown to the player, presented as a 3D primitive shape. The game asks the player to look at the item and describe it to the creature using gestures.

1 Unity Game Engine: https://unity3d.com

The game starts to record a gesture as soon as the player moves her hands, until either the time (30 seconds) runs out or the player places her hands in the rest position. A timer is displayed to indicate the amount of time left to perform the gesture, but the timer is shown only after 10 seconds of gesturing, so that the player might not feel like she is required to use all of the available time. Afterwards, a virtual character will repeat the gesture and the player can choose to acknowledge, or reject and perform the gesture again. This mechanism is basically providing feedback to the player on how the gesture looks like when performed through a virtual character and provides an 'undo' mechanism, a way for the player to correct a mistake.


***Figure 4**: Game screen captures of map (left) and part trading (right)*

After four shapes have been described, the player must venture out into the world to retrieve the missing parts. The game world is divided into 12 discrete zones, each designed as a different location in a sci-fi world. The missing parts are in possession of different creatures like the one encountered before, each at a different zone. This *virtual character* will perform a gesture, randomly picked from the pool. The player has to choose between four shapes, which shape best corresponds with the gesture she sees performed. There is also an "I don't know" option that triggers the avatar to perform a different gesture, while still referring to the same gesture class. The player has an incentive to do her best when labelling a gesture, because if she answered wrong, the character will leave and she will have to search for it at another location. A 'correct' (meaning, corresponding to the original ground truth) answer leads to the creature giving the player the ship part he was holding and bringing the player one step closer to completion of the game. Screen captures of this and the in-game map are shown by Figure 4.

## III.b Challenges and improvements

One interesting problem that we faced was how to signal the player that she can start gesturing and how to detect when a gesture is completed. Other experiments [44] use a traffic light like system. Our system asks the player to keep her arms alongside her body (resting position). Then a dialog box instructs her to start to gesture, however the system actually only starts to record when the hands leave the resting position. Recording ends when the hands enter resting position (for 1000ms) or when time runs out.

During development Bartertown underwent several iterations, based on user feedback. Earlier versions had a player hold out her arm for a determined amount of time to select a zone to visit or select a gesture class to label a gesture. Preliminary test sessions revealed that this was too cumbersome, so an interaction method was designed where the player just has to move her arm close her hand to confirm

the chosen option. Also, players had to wave one hand to select the “don't know” option in earlier versions. This turned out to be physically straining for users, so this too was changed. Players now have to make the so called 'scissor' gesture with two hands. This gesture involves holding the index and middle finger out, tucking the other fingers in and holding the arms away from the waist in an angle of around 30°. See Figure 5. This gesture is, however arguably arbitrary, not straining for the user, easy to detect by Kinect and unlikely to trigger any false positives.

***Figure** 5: The 'scissors' gesture, demonstrated by the game's protagonist*

Another notable area that underwent improvement were the in-game information texts displayed in dialog boxes. In the first iteration the game instructions had a considerable amount of non-essential, story related elements mixed in. While such flavour might help with immersion, players felt often confused and unsure what was expected of them. The text was thus changed to be less ambiguous. Numerous other small improvements were made, primarily to improve user interaction.

### III.c Data quality

Many of the common validation techniques rely on two players playing at the same time. Part of our challenge was to create a single-player environment. To help ensure data quality, the players engage in a form of multilevel review, as described by Quinn and Bederson [16]. In such a set-up, players label gestures that others playing before them have provided to the system. In fact, players may also label their own gestures, but this is slightly obfuscated by the fact that the gesture is in fact performed by a virtual agent.

# V. RESULTS

Data acquisition was conducted in April/May 2015. Participants were briefly instructed on how to interact with the game and were informed of Kinect's limited finger recognition capabilities. The ulterior motive of the game was explained to the participants. During the experiment, participants stood in front of a display, at a distance of about 2 meters. A play session lasted on average around 20 minutes.

## V.a Participants

During the experiment, 36 participants have played the game. Each player performed 4 gestures and labelled on average 11 gestures. The minimum amount of gestures a participant had to label to complete the game is 8, but this number is always higher because of incorrect labellings. Table 1 summarises some key characteristics of participants.

| | |
|---|---|
| gender | 20 male / 16 female |
| handedness | 32 right / 4 left |
| age | min 22, max 40, mean 28 |

***Table 1**: Participant characteristics*

## V.b Data

Table 2 demonstrates an aggregation on collected meta-data, clustered by intended gesture class.

| gestures performed | | gestures labelled | | | | | | | labelled as (confusion matrix) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| gesture class | n | correct | | incorrect | | don't know | | total | sph | cub | cyli | pyra | trap | hem | hou | octa |
| | | n | % | n | % | n | % | | | | | | | | | |
| sphere | 19 | 34 | 82.93 | 4 | 9.76 | 5 | 12.20 | 41 | 34 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| cube | 19 | 40 | 70.18 | 11 | 19.30 | 9 | 15.79 | 57 | 2 | 40 | 3 | 0 | 2 | 0 | 3 | 1 |
| cylinder | 18 | 30 | 46.15 | 26 | 40.00 | 9 | 13.85 | 65 | 1 | 6 | 30 | 3 | 8 | 2 | 3 | 3 |
| pyramid | 17 | 39 | 72.22 | 11 | 20.37 | 7 | 12.96 | 54 | 0 | 0 | 3 | 39 | 1 | 1 | 0 | 6 |
| trapezoid | 18 | 47 | 83.93 | 4 | 7.14 | 7 | 12.50 | 56 | 1 | 0 | 1 | 1 | 47 | 1 | 0 | 0 |
| hemisphere | 18 | 34 | 66.67 | 11 | 21.57 | 9 | 17.65 | 51 | 1 | 1 | 2 | 2 | 2 | 34 | 1 | 2 |
| house | 17 | 26 | 63.41 | 12 | 29.27 | 3 | 7.32 | 41 | 1 | 1 | 0 | 3 | 2 | 3 | 26 | 2 |
| octahedron | 18 | 37 | 67.27 | 12 | 21.82 | 6 | 10.91 | 55 | 1 | 0 | 0 | 1 | 5 | 3 | 2 | 37 |
| total | 144 | 287 | 68.33 | 91 | 21.67 | 55 | 13.10 | 420 | 41 | 49 | 40 | 50 | 67 | 45 | 35 | 51 |

***Table 2**: Aggregated gesturing and labelling data, per intended gesture class. A confusion matrix shows correct (green) and incorrect (red) labellings of performed gestures.*

Because the gesture classes for labelling are chosen at random, gesture labellings per gesture class fluctuate around the average of 52.5. It is clear to see that the *cylinder* shape was most difficult to label, with an error rate of 40% almost doubling the average (21.67) as well as the highest “don't know” score (13.85). It was most often confused with *cube* and *trapezoid*. *House* also has a rather high error rate (29.27).

*Correctness* is calculated for each gesture instance as the fraction of correct labelling (unknown responses are excluded from this calculation). We have calculated the *correctness* for each gesture instance in the dataset. Figure 6 shows a boxplot depicting lowest value, first quartile, median and third quartile. The mean *correctness* of the entire dataset = 0.77 (SD: 0.329).

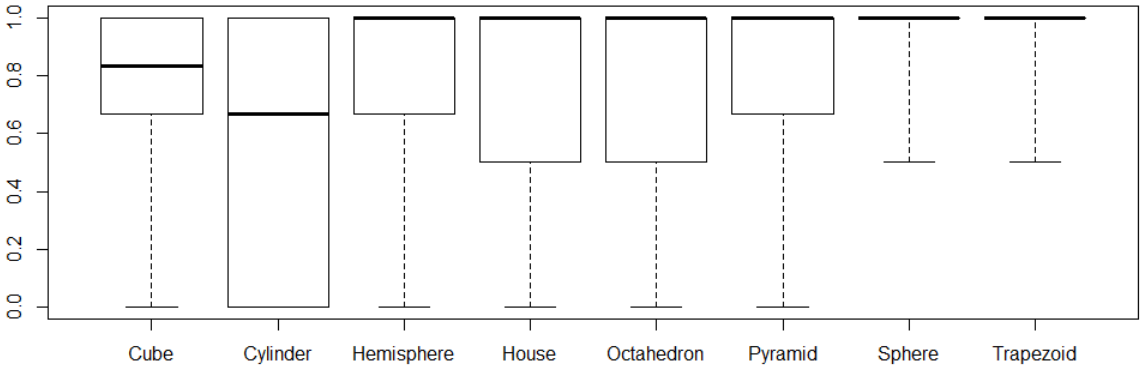


***Figure 6**: Correctness distribution of gesture instances, grouped by gesture class.*

## V.c Observations and questionnaire

Participants seemed to have little or no trouble with the navigation method. In fact, the user interaction was positively received, with participants remarking on the accuracy with which Kinect recognised their actions. However, they were disappointed to learn about the very limited manner in which the Kinect deals with hand positions. The hands have a lot of expressive power, that some participants also wanted to see reflected in the virtual avatar. Most participants used the 'drawing' strategy to express the shapes. A minority tried to use their posture. However, in all cases those participants that try to use postures experience that some shapes are impossible to convey in this manner, and eventually resort to drawing. Most participants drew 2 dimensional shapes, a minority tried to express the third dimension as well.

The majority of the participants did not seem inclined to choose the 'I don't know' option during gesture labelling. When they were in doubt, they seemed to prefer to take a guess rather than choose the 'I don't know' option. If this is because this option was perhaps less obvious to the participants or they simply preferred to take a gamble is not entirely sure. Some participants expected some interaction during the intro cinematic (they were for example trying to fly the plane with their hands), even after they had been instructed that such actions would have no effect. Most participants seemed to have a good time and made positive remarks about the game play, graphics and animations.

After completing the game, participants were asked to fill out a survey (see X.c Questionnaire results). Summarising, the majority of participants (60.71%) said to have little trouble selecting the correct shape when the virtual avatar performed a gesture. 70% also recognised their own movements when reflected through the avatar. 46.67% said they enjoyed the game and would play again. 26.67% said their prime motivation for playing would be contributing to research and another 16.67% said they prefer other types of games. 90% of participants agreed that the visual aesthetics made a positive contribution to their play experience.

# VI. GUIDELINES

This section describes some brief guidelines that have been developed to help researchers during the task of using human computation games for creating gesture datasets.

- **Consider dataset first**
  The project should always be centred around the dataset, so this has to be designed first. A gesture vocabulary should be drawn up, each class should sufficiently differ from others to make unambiguous ground-truth labelling possible, yet also have a degree of uniformity. The guidelines by Ruffieux [32] will prove useful here. We advise to be mindful of the fact that by using a game to collect gestures one introduces a new, perhaps unnatural context. Players could be primed by playing the game and not perform natural gestures.

- **Define technical requirements**
  The following technical aspects are relevant:
  1) What sensor equipment (and corresponding software) will be employed. We opted for a Kinect v2 for its full body tracking capabilities, affordability and available ease of use SDK. However, depending on your scenario, it could be useful to consider other alternatives. Leap Motion[2] for example, is far better suited for detecting hand positions. However, the use of specialised equipment does make deployment on a large scale an issue.
  2) How the data will be stored. The most common ways of presenting gesture corpora is through motion video and comma separated text files containing long lists of information on limb position. We however, have chosen a formal relational database because it makes access to the data in the game environment easier.
  3) What platform / programming language for the game. This depends on a plethora of factors, price and licence, availability of assets and support, target platform, knowledge and preference of programming language, etc. We have found Unity3D to work very well.

- **Design game rules**
  A number of important aspects about the game design have to be considered. Herein lies the true challenge of any human computation game, finding a game mechanic that on the one hand provides the desired data with sufficient data quality and on the other is fun enough to motivate people to play it. There has been substantial research into game design for human computation tasks [16]. One must choose whether the game is to be played solo or real-time multi-player. Mind that because humans are imperfect in recognising gestures it might be required to label a gesture multiple times. The game design should encourage 'correct' actions and discourage wrong actions. We also encourage to offer the players an “I don't know” option.

2 Leap Motion: https://www.leapmotion.com

- **Choose representation of gestures**
  If gesture performances are to be independently labelled by other players, it is important to consider the notion that players might (inadvertently or not) communicate (part of) the message not just through gestures but external objects (e.g. a ball, a written note). To prevent that, we have chosen to remove all unnecessary context and abstracted the gestures through a virtual avatar. Other anti-cheating measures are also possible.

- **Provide feedback**
  It is imperative that players receive immediate and unambiguous feedback of their actions and the reactions of the system. Failing to do so will leave players confused and frustrated. Conforming to common game design literature, players should have a clear goal and be able to monitor their progress to that goal [45].

- **User interaction and aesthetics**
  User interaction design (UID) is vitally important in any computer game. When UID is felt lacking or cumbersome, players could accidentally trigger an unintended action. What's more, an unresponsive or unintuitive user interface could frustrate players and cause them to stop playing. UID also encompasses the dialogue dimension. We recommend in-game information dialogs to be as unambiguous as possible, even if this is at the cost of immersion into the game. However, providing a background story and aesthetics through interesting visual art and sounds is a very strong factor in the game experience. For both of these aspects, it is important to consider the target demographic.

# VII. Discussion

## VII.a Conclusion

In this work, we have introduced *Bartertown*, a Kinect based human computation game for the creation of an iconic gesture dataset. Special effort has gone out to take the human computation paradigm to the next logical level, using story and visual aesthetics to deliver an immersive and enjoyable experience. To create a dataset that is usable for researchers, data quality is of the essence. To ensure data quality, the game features a self-validating feature where earlier gesture recordings are evaluated by other players.
The results of our prototype experiment are encouraging, most of the gestures performed are positively labelled to the correct ground truth. Furthermore, players enjoyed the experience and reported willingness to play again. We have collected a sizeable dataset of annotated gestures (n = 144). It is available to download online[3]. We have also drawn up a number of general guidelines that can be used heuristically when considering a human computation game approach for collecting gesture corpora.

## VII.b Future work

Future improvements to *Bartertown* could be to add RGB video recording capabilities. Within the context of this graduation project, we did not have sufficient resources to implement this. Recorded full motion video can be of great value to researchers. However, it would have to be a feature that players can opt-out of, for privacy reasons. Some gestures proved to be too difficult and thus frustrating for players. To improve the fun of the game it could be considered to have an arbitrary cut-off point, a number of mistakes for one gesture instance after it drops off from the pool of available gestures.

Future additions to a human computation approach to creating a gesture dataset could entail some way to cluster variations of a gesture class, so as to group the different gestures based on some class characteristics (e.g. all posturing gestures). Another potentially very interesting addition could be the addition of temporal ground-truthing, (i.e. differentiating between *pre-stroke*, *nucleus* and *post-stroke* stages of a gesture) Research on temporal ground truth segmentation using crowdsourcing platforms [46] has met with some promising results. We have not added these features to the prototype discussed in this paper because of added complexity.

Participants were limited in their freedom of expression because of Kinect's native hand tracking capabilities are rather sub-par at the time of the experiment. However, some promising results have been found by using model based tracking methods [47]. Future researchers interested in doing complex gesture tracking using Kinect would do well to consider this or other non-standard approaches.

Bartle's taxonomy of player types [48] describes four archetypes: *killers*, *achievers*, *socialisers* and *explorers*. Obviously, our game mainly appeals to the latter. It could be very interesting to expand the game to also offer content for other types of players. One could think about multi-play capabilities, communication, integration with social networks and many of the tried and true methods of gamification, e.g.: leaderboards, achievements, virtual currencies etc.

3 Download the dataset at: http://mediatechnology.leiden.edu/openaccess/bartertown

The scope of this work was limited to a small number of 3D primitives for our vocabulary. Future games with a purpose could also be employed to expand these classes with more shapes, or other classes such as abstract or metaphoric concepts, actions, affective states, etc.

Although one of the biggest advantages of using a human computation game is the potential for a large amount of participants we still conducted our experiment in a lab setting. The requirement of a depth camera made distributed deployment difficult, so we consider this implementation a proof of concept. However, recently we have noticed a surge of consumer grade laptops and tablets outfitted with a webcam with depth sensing capabilities[4]. This is a promising development indeed for the future of human computation games for gesture corpora.

4 Intel® RealSense™: http://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

# VIII. ACKNOWLEDGEMENTS

We would like to gratefully thank Maarten Lamers, Amir Sadeghipour, Myriam Traub, Remi Alkemade, Alan Cienki and Jasper Kok for their contribution.

# IX. REFERENCES

[1] V. I. Pavlovic, S. Member, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[2] H. Park, E. Kim, S. Jang, and S. Park, "HMM-based gesture recognition for robot control," *Pattern Recognition and Image Analysis, Pt 1, Proceedings*, vol. 3522, pp. 607–614, 2005.

[3] H. Kang, C. Woo Lee, and K. Jung, "Recognition-based gesture spotting in video games," *Pattern Recognition Letters*, vol. 25, no. 15, pp. 1701–1714, Nov. 2004.

[4] H. Narahara and T. Maeno, "Factors of Gestures of Robots for Smooth Communication with Humans," *The Proceedings of First International Conference on Robot Communication and Coordination*, p. 44, 2007.

[5] C. Huang and B. Mutlu, "Modeling and Evaluating Narrative Gestures for Humanlike Robots.," *Robotics: Science and Systems*, pp. 57–64, 2013.

[6] S. W. Cook and S. Goldin-Meadow, "The Role of Gesture in Learning: Do Children Use Their Hands to Change Their Minds?," *Journal of Cognition and Development*, vol. 7, no. 2, pp. 211–232, Apr. 2006.

[7] W.-M. Roth and D. Lawless, "Scientific investigations, metaphorical gestures, and the emergence of abstract scientific concepts," *Learning and Instruction*, vol. 12, no. 3, pp. 285–304, Jun. 2002.

[8] V. Richmond, "Teacher nonverbal immediacy: Uses and outcomes.," in *Communication for Teachers*, J. L. Chesebro and J. C. McCroskey, Eds. Boston: Allyn & Bacon., 2002, pp. 65–82.

[9] S. Goldin-Meadow, "Talking and Thinking With Our Hands," *Current Directions in Psychological Science*, vol. 15, no. 1, pp. 34–39, Feb. 2006.

[10] K. Bergmann and S. Kopp, "GNetIc – Using Bayesian Decision Networks for Iconic Gesture Generation," in *Intelligent virtual agents*, vol. 5773, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 76–89.

[11] S. Mitra, S. Member, and T. Acharya, "Gesture Recognition : A Survey," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311–324, 2007.

[12] J. Eisenstein and R. Davis, "Visual and Linguistic Information in Gesture Classification Categories and Subject Descriptors," *Proceedings of the 6th International Conference on Multimodal Interfaces*, pp. 113–120.

[13] T. Darrell and A. Pentland, "Active Gesture Recognition using Partially Observable Markov Decision Processes," in *International Conference on Pattern Recognition*, 1996, pp. 984–988.

[14] K. Murakami and H. Taguchi, "Gesture Recognition using Recurring Neural Networks," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1991, pp. 237–242.

[15] M. A. Priesters and I. Mittelberg, "Individual differences in speakers' gesture spaces: Multi-angle views from a motion-capture study," in *Proceedings of the Tilburg Gesture Research Meeting (TiGeR 2013)*, 2013, pp. 1–4.

[16] A. J. Quinn and B. B. Bederson, "Human Computation: A Survey and Taxonomy of a Growing Field," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1403–1412, 2011.

[17] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, Feb. 2011.

[18] P. Hsueh, P. Melville, and V. Sindhwani, "Data Quality from Crowdsourcing : A Study of Annotation Selection Criteria," in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 2009, no. June, pp. 27–35.

[19] L. Von Ahn and M. Reiter, "Human Computation," School of computer Science, Carnegie Mellon University, 2005.

[20] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, 2004, vol. 6, no. 1, pp. 319–326.

[21] L. Von Ahn, R. Liu, and M. Blum, "Peekaboom : A Game for Locating Objects in Images," in *Proceedings of the SIGCHI conference on Human*

*Factors in computing systems*, 2006, pp. 55–64.

[22] O. Chrons and S. Sundell, "Digitalkoot : Making Old Archives Accessible Using Crowdsourcing," in *Artificial Intelligence*, 2011, pp. 20–25.

[23] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg, "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, Sep. 2008.

[24] R. Alkemade, "Hands-On 3D Design in Virtual Reality," Master's Thesis for the Media Technology programme, Leiden University, 2015.

[25] P. Ekman and W. V. Friesen, "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding," in *SEMIOTICA*, A. Kendon, Ed. DE GRUYTER, 1969, pp. 49–48.

[26] D. Efron and S. van Veen, *Gesture, race and culture*. Crown Press, 1972.

[27] D. McNeil, "Gestures of the Concrete," in *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992, pp. 105–133.

[28] D. McNeill, "Guide to Gesture Classification, Transcription and Distribution," in *Hand and Mind: What Gestures Reveal about Thought*, University of Chicago Press, 1992, pp. 75–103.

[29] D. Casasanto and S. Lozano, "The meaning of metaphorical gestures," *Metaphor and Gesture.*, 2007.

[30] U. Hadar and L. Pinchas-Zamir, "The Semantic Specificity of Gesture: Implications for Gesture Classification and Function," *Journal of Language and Social Psychology*, vol. 23, no. 2, pp. 204–214, Jun. 2004.

[31] R. M. Krauss, P. Morrel-Samuels, and C. Colasante, "Do conversational hand gestures communicate?," *Journal of personality and social psychology*, vol. 61, no. 5, pp. 743–54, Nov. 1991.

[32] S. Ruffieux and D. Lalanne, "A Survey of Datasets for Human Gesture Recognition," *Lecture Notes in Computer Science*, vol. 8511 LNCS, no. PART 2, pp. 337–348, 2014.

[33] A. Sadeghipour, L.-P. Morency, and S. Kopp, "Gesture-based object recognition using histograms of guiding strokes," in *British Machine Vision Conference*, 2012.

[34] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, p. 1737, 2012.

[35] E. L. M. Law, L. Von Ahn, R. B. Dannenberg, and M. Crawford, "TagATune : A Game for Music and Sound Annotation," *In International Conference on Music Information Retrieval*, pp. 361–364, 2007.

[36] B. Borsboom, "Guess Who ?: A game to crowdsource the labeling of affective facial expressions is comparable to expert ratings .," Master's Thesis for the Media Technology programme, Leiden University, 2012.

[37] R. Speer, C. Havasi, and H. Surana, "Using Verbosity : Common Sense Data from Games with a Purpose," in *Florida Artificial Intelligence Research Society Conference*, 2010, vol. 2000, pp. 104–109.

[38] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players, "Predicting protein structures with a multiplayer online game.," *Nature*, vol. 466, no. 7307, pp. 756–60, Aug. 2010.

[39] A. L. Robinson, "SpotOn NYC: Communication and the brain – A Game To Map The Brain," *nature.com*, 2013. [Online]. Available: http://www.nature.com/spoton/2013/03/spoton-nyc-communication-and-the-brain-a-game-to-map-the-brain/. [Accessed: 16-Jul-2014].

[40] A. Lieberoth, M. K. Pedersen, A. C. Marin, T. Planke, and J. Sherson, "Getting humans to do quantum optimization: user acquisition, engagement and early results from the citizen cyberscience game Quantum Moves," *Human Computation*, vol. 1, no. 1, 2014.

[41] T. Sowa and I. Wachsmuth, "Interpretation of shape-related iconic gestures in virtual environments," *Gesture and sign language in human-computer interaction*, vol. 1, pp. 21–33, 2002.

[42] J. Šimko, M. Tvarožek, and M. Bieliková, "Human computation: Image metadata acquisition based on a single-player annotation game," *International Journal of Human Computer Studies*, vol. 71, no. 10, pp. 933–945, 2013.

[43] M. Krause, M. Wittstock, and R. Malaka, "Frontiers of a Paradigm – Exploring Human Computation with Digital Games," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010, pp. 22–25.

[44] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled, "Gesture Recognition Corpora and Tools: A Scripted Ground Truthing Method," *Computer Vision and Image Understanding*, Aug. 2014.

[45] J. Schell, *The Art of Game Design: A book of lenses*. CRC Press, 2014.

[46] M. Burlick, O. Koteoglou, L. Karydas, and G. Kamberov, "Leveraging Crowdsourced Data for Creating Temporal Segmentation Ground Truths of Subjective Tasks," *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 743–750, Jun. 2013.

[47] I. Oikonomidis, N. Kyriazis, and A. a Argyros, "Efficient Model-based 3D Tracking of Hand Articulations using Kinect," *22nd British Machine Vision Conference*, pp. 1–11, 2011.

[48] R. Bartle, "Hearts, clubs, diamonds, spades: Players who suit MUDs," *Journal of MUD research*, vol. 1, no. 1, p. 19, 1996.

# X. APPENDICES

## X.c Questionnaire results

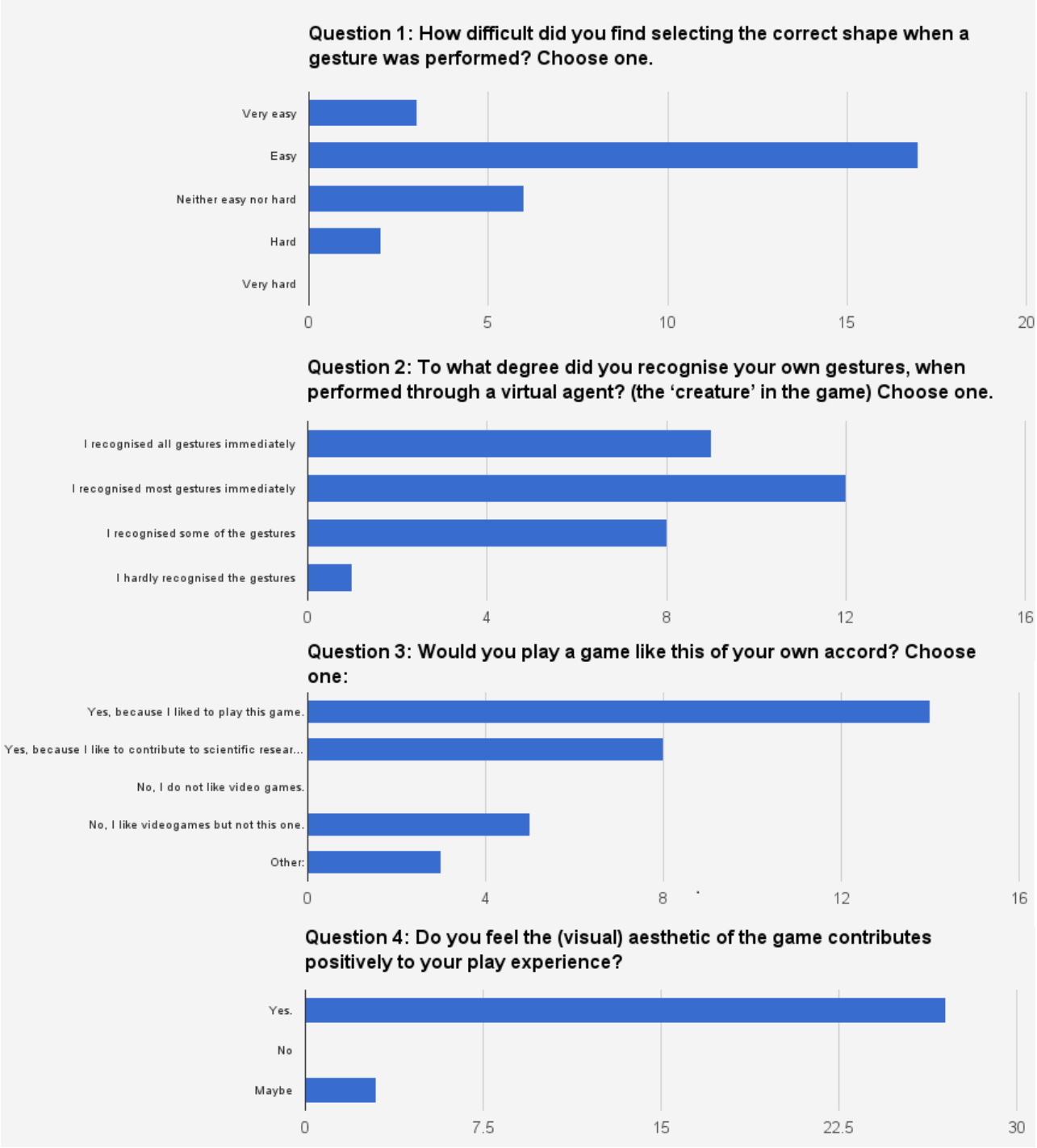


***Figure 7**: Questionnaire results*