Evaluating Digital Avatars in VR - A Systematic Approach to Quantify the Uncanny Valley Effect

Hakan Arda
Faculty of Computer Science and Business Information Systems
Technical University of Applied Sciences Würzburg-Schweinfurt
Würzburg, Germany
e-mail: hakan.arda@thws.de

Andreas Henneberger
Faculty of Computer Science and Business Information Systems
Technical University of Applied Sciences Würzburg-Schweinfurt
Würzburg, Germany
e-mail: andreas.henneberger@study.thws.de

***Abstract*—Virtual reality has undergone drastic developments in recent years. At the forefront are almost perfectly rendered real-life objects that can hardly be distinguished from the original. However, while the improvements in the realism of the environment significantly increase the user's immersion in the world, the increase in the realism of human-like avatars seems to have the opposite effect. Extremely realistic depictions of people in computer games, films, or other immersive applications often evoke negative feelings and can thus lead to a break in immersion. This emotional break is illustrated by looking at the uncanny valley curve. In this work, we have tried to develop a way to evaluate human-like avatars according to the uncanny valley curve and thus to determine more precisely where the discomfort comes from. To do this, we created a database of over 200 images of avatars and used studies to determine various precise characteristics that make these avatars like humans. In addition, we were able to evaluate the approach of this work with a pilot study and thus offer a possibility for evaluations of avatars according to the uncanny valley for future research.**

Keywords—Uncanny valley; avatars; human-likeness; database; virtual reality.

I. INTRODUCTION

The Uncanny Valley effect is a psychological phenomenon that describes the unease or discomfort people experience when encountering human-like entities that are almost, but not quite, convincingly realistic. This emotional break is illustrated by looking at the uncanny valley curve. In Figure 1, you can see a slight increase until the graph reaches around the 50 percent mark and then drops rapidly. This drop is known as the uncanny valley. The term “Uncanny Valley” was coined by robotics professor Masahiro Mori in 1970 [1]. The concept suggests that as the appearance or behavior of humanoid entities becomes increasingly close to human-like, there is a point at which they elicit a strong negative emotional response before eventually becoming indistinguishable from real humans.

There have already been many far-reaching attempts to investigate the effects of the human likeness of robots on people's emotional reaction [2–8]. One example of this is the work of Kim et al. [6], who used the open-source Anthropomorphic RoBOT (ABOT) database to analyze the human similarity of 251 robots. They asked a group of 150 participants to rate images of robots from the ABOT database according to their human likeness and uncanny valley factor. With the results of this survey, they have found evidence of Mori's uncanny valley [1]. This valley was evident in participants' perceived uncanniness of 251 robots that varied widely in terms of the range and characteristics of human likeness. They also found evidence of another, second valley of uncanniness in robots that showed a moderately weak resemblance to humans.

The developers of the ABOT [3] database took a similar approach in their study, providing a basis for research in this area. The researchers found that the human-like appearance of robots can be divided into three dimensions of human-like appearance: the robots' surface features (e.g., skin, hair, clothing), the main components of the robots' body manipulators (e.g., torso, arms, legs) and the robots' facial features (e.g., eyes, mouth, face) [6]. These results suggest that the overall perception of the physical human-likeness of robots and its relationship to emotional reactions to the robots can be explained by different constellations of the three human-like appearance dimensions. If the hypothesized uncanny valley phenomenon could be understood at the level of specific human-like appearances, this could also lead to the improvement of virtual avatars.

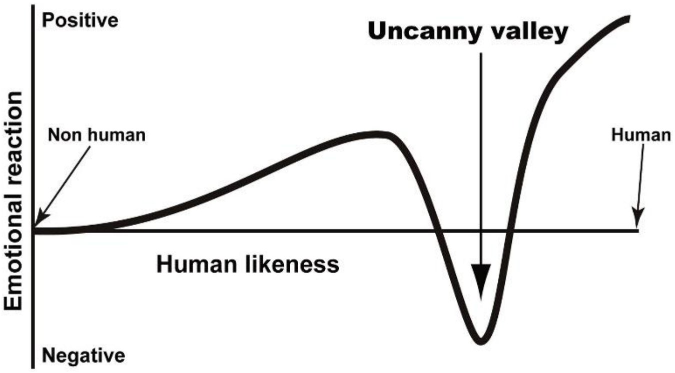

Figure 1. Graphical representation of the uncanny valley. Retrieved from [9].

This approach of using a large database, such as ABOT, to evaluate different human-like robots for their uncanny valley factors does not exist in relation to digital avatars currently. One

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-163-3

reason for this could be the absence of such a database and the associated basis for the resulting research. Another reason is that there is currently no systematic, evidence-based approach for categorizing avatars into a continuum of perceived human-likeness. Consequently, researchers and designers are usually forced to rely on heuristics and intuitions when it comes to selecting human-like avatars for studies or developing human-like features in avatar design. This approach faces several problems. Firstly, there is currently no quantitative system to describe the degree of human likeness in different avatars, which makes it difficult to compare research results between different studies. Therefore, a precise scale is needed to compare different avatars on a common scale and allow researchers to replicate their results with their avatars.

Secondly, even when researchers manage to quantitatively assess the impression people have of an avatar's appearance, they usually treat the concept of “human likeness” as one-dimensional. However, human likeness can be expressed in different ways. For example, through gestures and facial expressions or, more generally, through the mere presence of arms and legs. Humanity has many different characteristics and therefore also different features that need to be considered.

Thirdly, the effects of the appearance of avatars on different types of avatars must be considered in the investigations. While it may be of practical advantage to limit yourself to a certain type of avatar, such as simplified human likenesses like the ones Meta uses in her Horizon. However, these constraints can mean that certain minor differences between avatars are lost. Consequently, the psychological insights from such work may not generalize to a diverse range of avatars.

In addition, previous studies have extensively explored methods to overcome the uncanny valley in character design. For example, the work by Schwind et al. examined how atypical features (strong deviations from the human norm) for high levels of realism cause negative sensations in humans and animals [10] [11]. The negative effects of atypical features, such as unnaturally large eyes or human emotions in realistic animals, are stronger for more realistic characters, than for characters with reduced realism. Consistently, rendered realism can reduce the negative effects of atypicality and increase affinity, as shown by the first peak in the uncanny valley. Therefore, it is important to avoid combining realistic renderings and detailed textures of skin or eyes with non-human-like features. At high levels of realism, atypical features can cause the uncanny valley. Another possibility is to avoid so-called “dead eyes”. A virtual character's eyes are crucial in determining its realism. The work of Schwind et al., which used eye tracking, found that users first fixate on the eyes before assessing other features. This is consistent with previous research showing that the realism and inconsistencies of human characters are primarily judged based on the realism of their eyes [12]. This also explains why skin makeup does not affect animacy, unlike atypical eyes or the eyes of a deceased person. The symptom of “dead eyes” can make artificial characters feel eerie and strange. The eyes communicate intentions, behavior, and well-being, which are essential for assessing and creating affinity for the depiction. It is important to clarify this and many other problems regarding the appearance of the avatars beforehand in order to increase user acceptance. Uncanniness could distract the user from the actual idea and thus reduce acceptance [11].

To solve these problems and to offer a way to conduct more systematic, general, and repeatable research on virtual human-like avatars, we developed a database for avatars based on the findings of Phillips et al. and thus continues the research of the uncanny valley. This paper begins by highlighting the problems that arise when there is no fixed way to evaluate the concept of the uncanny valley. Section 2 presents the methodologies employed to construct the database of avatars, the instructions for the human-likeness and uncanny valley questions, and the specific details of the survey. Section 3 presents the results of the survey and Section 4 interprets the results accordingly. Finally, Section 5 summarizes the results of the collected work.

II. Method

A. The Avatar Database

Similar to the ABOT database by Phillips et al., the avatar database pursues three goals. Firstly, it offers an overview of the broad landscape of different human-like avatars. Secondly, the avatar database provides standardized images of human-like avatars and a growing dataset of people's perceptions of these avatars, both of which will be made public for further research in the future. Thirdly, the Avatar Database can help us better understand what makes an avatar appear human. To begin, we will discuss the development of the database. Then, two empirical studies will identify various dimensions of avatar appearance and determine which of these dimensions contribute to the perception of overall human-likeness of avatars. Subsequently, a further empirical study on the validation of the database is presented and a possibility for future research is introduced.

B. The Development of the Avatar Database

To create a comprehensive database of human-like avatar images, we searched for as many avatars as possible with the required human-like appearance characteristics. The avatars were identified from various sources, such as game characters, technology-oriented media, companies and university websites, online communities and discussion forums, and general Google searches. To identify avatars that had not yet received significant media attention, we also created our own avatars based on real humans and fictional characters, using various modular systems and scanning tools. Between January 2024 and April 2024, we created an initial collection of over 200 images of human-like avatars, each with one or more different characteristics.

Next, we reviewed the collection of avatars and removed those that were already represented in the same or similar enough form, for example avatars that only differed in clothing but not in body. In addition, avatars that are too similar to animals have also been removed because this study explicitly looks at human avatars. Each image also had to fulfil a certain standard to be included in the collection: No obstacles, no motion blur, no groups pictures, in color, and the entire body (With feet, hands, and head). In addition, all pictures were taken one more time in a close-up of the face. This allowed the participants to examine the entire body and then explicitly the face for the individual features. Any image that did not fulfil this standard was either not included in the collection or removed accordingly. For characters that were important for the collection, such as avatars that reflect a well-known personality, we created a suitable avatar ourselves. We also included avatars

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-163-3

based on images created by artificial intelligence (Midjourney) and labelled them for later investigation. Based on this approach, we have created a collection of 200 avatars, with the corresponding source for downloading the avatars.

It was important to us that we cover as many different types as possible with the large number of avatars, such as cartoonish, stylized, realistic or minimalist. In addition, it should be possible to find an avatar that is reduced to the desired characteristics with the help of filters. This is a similar concept to searching for robots in the ABOT database. With the difference that here the avatars can be downloaded with the help of the source and used for further purposes. Some of the characters were also tested with the help of the unity engine and put to the test for suitable images. This was especially the case with avatars that we developed ourselves for this collection.

The images in the database were sorted and edited to ensure uniform recognition. This was done to ensure consistency in the images. Avatars were photographed in a frontal and neutral position with a neutral facial expression whenever possible. For avatars where this was not possible, and the model was available for free download we rendered them again in the unity engine and photographed accordingly. Finally, the images were cropped to just the avatars with a white background using photoshop and tagged with different tags for better analysis. Here, for example, attention was paid to the recognizable gender, potential age, art style and source. In Figure 2 you can see all the avatars used for this study.

C. Measuring the human-likeness

To accurately determine the degree of humanity in avatars, the individual avatars must be evaluated according to clearly defined characteristics. Because we are dealing here with human-like avatars and not anthropomorphic robots, we are unfortunately unable to use the results of the work by Phillip et al. [3] But, we can use a similar approach to determine the characteristics that we will use later. We rely on a bottom-up, feature-based approach and base our expectations on the results of the work of Phillip et al. Our goal with this approach was to define the individual features that constituted humanity in avatars and then bundle them together.

To determine the appearance characteristics of our avatars, we have created a collection of possible characteristics based on the work of Phillip et al. [3]. We then checked all the images for the respective characteristics and deleted any that did not fit. This included features that were rare, repeated, or confusing. As a result, we defined 16 characteristics that we used for our further procedure. In addition, like Phillip et al., we decided to contribute definitions for the features. We started with relevant definitions from the Oxford English dictionary [13] and adapted them according to our application. For example, we were able to retain the biological functions for features such as “mouth” because our test objects are not robots but human-like avatars. This resulted in a table of features and their definitions. These definitions served as a way for our participants to focus on certain characteristics when evaluating the avatars. Since they are human-like avatars and not robots, all characteristics are always present in some form. They only differ in their design. For example, each avatar has a “skin” that can be black, white, brown, et cetera. However, there are also avatars that do not have smooth white skin and are instead green with lots of dots. Therefore, the question here is not whether the respective characteristic is present, but rather to what degree it stands out. With the help of these characteristics, we were able to provide clear points of reference for the participants to evaluate the avatars.

TABLE 1. COLLECTION OF APPEARANCE FEATURES AND ASSOCIATED DEFINITIONS

Feature	***Definition***
Arm	The upper limb of the human body, or the part of the upper limb between the shoulder and the wrist.
Eye	The organ of sight. Either of the paired globular organs of sight in the head of humans.
Eyebrow	The (usually arched) line of short fine hair along the upper edge of each of a person’s eye sockets.
Eyelashes	The line of hairs fringing each edge of an eyelid, serving to help keep the eye free of dust or other extraneous matter.
Face	The front part of the head, from the forehead to the chin, and containing the eyes, nose, and mouth.
Finger	Each of the five slender jointed parts attached to either hand.
Genderedness	Features of appearance that can indicate biological sex, or the social categories of being male or female.
Hand	The terminal part of an arm, typically connected to the arm by a wrist. A hand is normally used for grasping, manipulating, or gesturing.
Head	The uppermost part of a body, typically connected to the torso by a neck. The head may contain facial features such as the mouth, eyes, or nose.
Head hair	A collection of threadlike filaments on the head.
Leg	The lower limb of the human body, or the part of the lower limb between the hip and the ankle.
Mouth	The orifice in the head of a human or other vertebrate through which food is ingested and vocal sounds emitted.
Nose	The part of the head or face in humans which lies above the mouth and contains the nostrils.
Skin	The layer of tissue forming the external covering of the body.

D. Measuring the uncanniness

For the question of how uncanny the avatars are, we used the study by Kim et al. [6] as a guide. They also conducted a large-scale study with the help of the ABOT database and, based on the results of Phillip et al. [3], conducted a follow-up study in

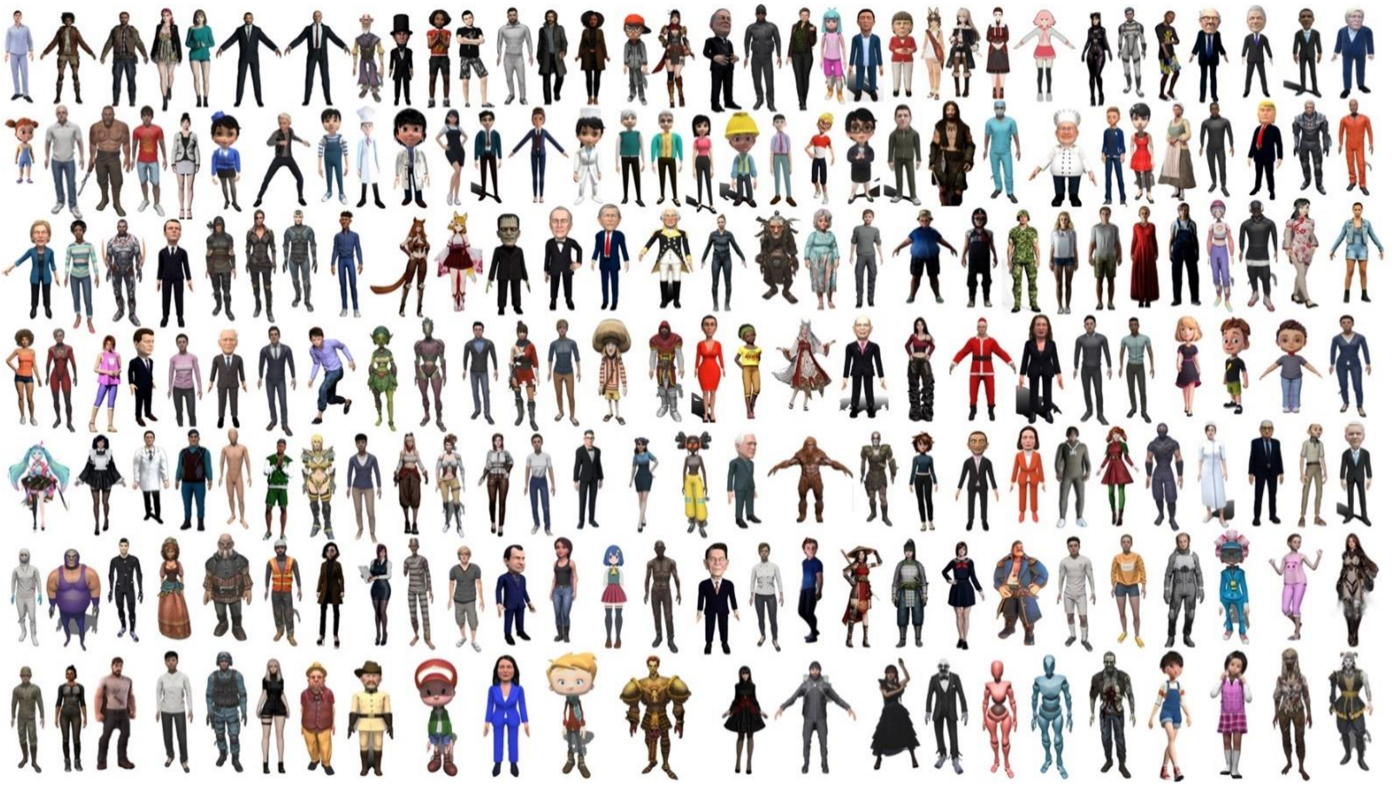

Figure 2. All 200 human-like VR avatars in the database.

which they compared the robots in terms of human-likeness and uncanniness. We also used the same definition of uncanniness to encourage participants to apply a standardized criterion in the evaluation. The participants were given the following definition, uncanniness is: “The characteristic of seeming mysterious, weird, uncomfortably strange or unfamiliar.” This definition was derived from the Oxford English dictionary's definition of uncanniness.

E. *Participants*

For the study, we recruited a total of 160 participants via Prolific crowdsourcing website. Data collected via crowdsourcing, websites such as Prolific is currently very much in vogue. This is mainly due to the fact that the data can be collected very easily and quickly and there are already studies showing that the data collected here can keep up with traditional methods in terms of quality [14]. Nevertheless, the data should also be checked for quality [15]. We have therefore decided to incorporate various quality checks into the data collection process. Firstly, all data sets with incorrect answers to six or more “catch trials” are removed. Secondly, we considered a lack of variation in ratings between participants as an indicator of inattention. Therefore, we removed data from participants whose ratings had a standard deviation of less than 10 (SD < 10) on a scale of 0 - 100. Finally, we compared each participant's ratings with the average of the remaining judgements in their group (between participants) by calculating the correlation between the individual judgements and the remaining judgements in their group. If this correlation between the individual participant's ratings and the group mean was less than 30, the participant's data were discarded as these individuals may have been performing a different assessment task to the group. After this quality check, the total number of participants was 143 (M Age = 20, SD Age = 10, 104 Male, 41 Female, 2 No Responses). This means that each avatar had a rating of 15 - 20 participants.

F. *Design and procedure*

The 200 avatars were divided random into four groups of 50 avatars each. Each group was also provided with 10 catch trials. This meant that each group had 60 avatars, which were rated by 20 participants. Because two different questions were asked in this study, one asking, “how human-like is the avatar?” and the other “how uncanny is the avatar?”, we asked each group twice. This gave us a total of 4 groups of 60 avatars per question.

The participant begins the survey with a brief introduction to the topic and a short briefing on how to complete the survey. This was followed by an example task on how the participant should rate the avatars. The same example avatar was used for each block. The participant sees two pictures of an avatar on their screen. On the left the entire body and on the right the profile picture with the face in focus. Below the pictures is the definition of the respective question. For the question about human likeness, the participant sees the various characteristics that make an avatar like humans and a slider from 0 to 100. Above the slider is the question “How similar to humans do you think this avatar is?” (0 - not at all like humans and 100 - very similar to humans). We used a similar method for the question of how uncanny you think the avatars are. You can see the same pictures and again a slider from 0 to 100 but this time with the definition about uncanniness and the question “How uncanny do you think this avatar is?” (0 - not uncanny at all and 100 - very uncanny). The participants are randomized into one of the respective groups and are only allowed to answer one question

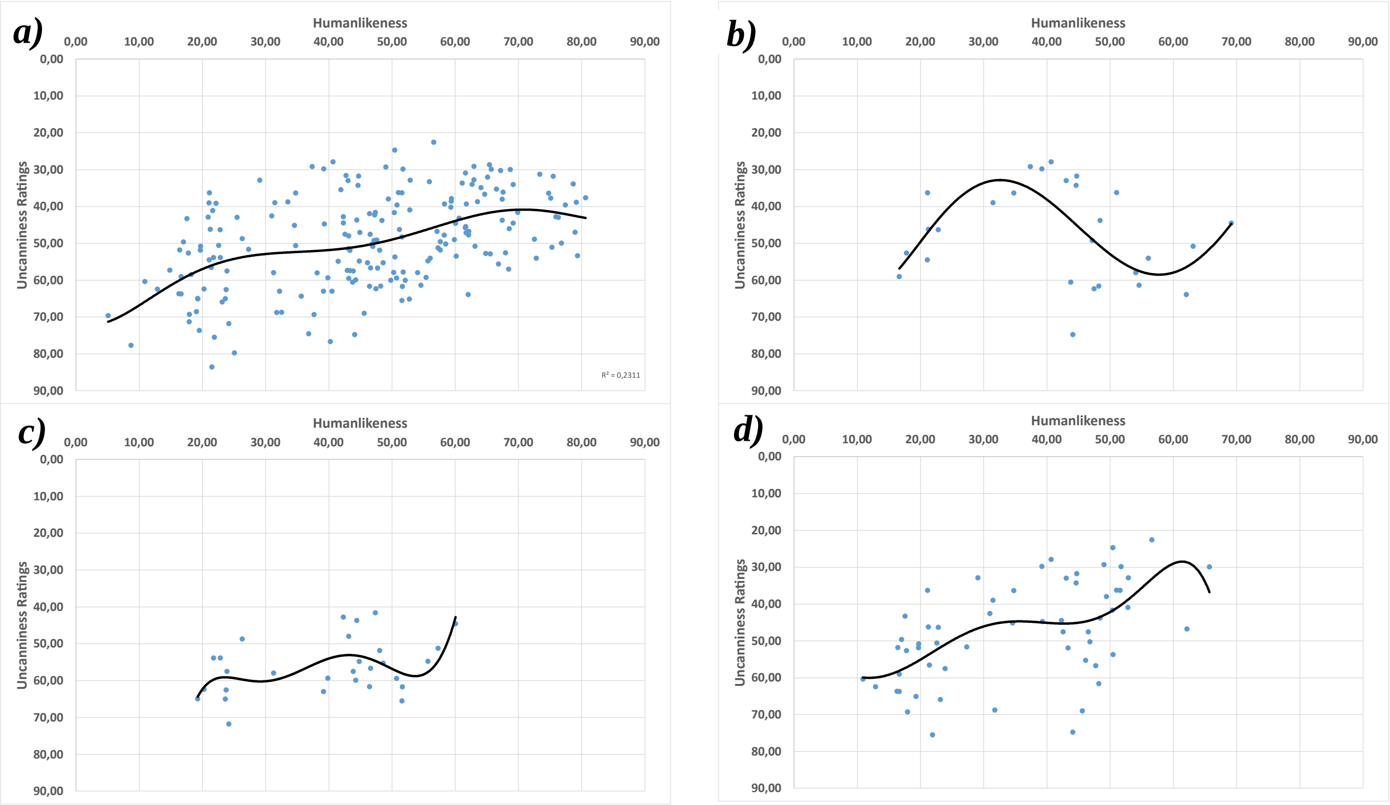

Figure 3. Four scatterplots a) Total scope of all avatars b) Only avatars representing children c) Avatars caricaturing a real-life person d) Avatars representing a cartoonized style. This Y axis has been inverted to represent the uncanny valley.

type. This is to prevent the questions from influencing each other. The catch trials are pictures of real people or objects that have also been randomized into the respective groups.

After half of the questions, the participants were given a 10-second break during which their attention was drawn to the definition and characteristics again. After judging all the images, participants were asked to complete a demographic questionnaire in which they were asked to indicate their age, gender, native language, level of education, previous knowledge of robotics and experience with virtual avatars. The entire study took approximately 5 minutes to complete, and participants received $1 as compensation for their participation.

III. Results

For the data analysis, all the results of the individual surveys were added together and an average for human likeness and uncanniness was calculated for each avatar. We then inserted these results into Microsoft Excel to generate various graphs. Looking at the first graph (Figure 3 a), no uncanny valley can be recognized. Instead, there is a linear gradient between the two factors uncanniness and human likeness, with uncanniness decreasing as human likeness increases. However, if one does not look at the entire amount of data and instead only at certain categories, such as only avatars that represent children, an uncanny valley is clearly recognizable. Just as in Mori's uncanny valley hypothesis [1], a large valley can be recognized between the moderately realistic and realistic avatars. When looking at other avatar categories, a slight uncanny valley can also be recognized. The other avatar categories, such as avatars that are based on a real person and represent them as a caricature, also have a slightly uncanny valley. The same applies to avatars that are not based on a real person but are depicted as a cartoon. based on a real person and represent them as a caricature, also have a slightly uncanny valley. The same applies to avatars that are not based on a real person but are depicted as a cartoon. To further investigate these results, we performed a polynomial mixed fit for the three different categories of. We determined the different coefficients of determination = r^2 for different polynomial mixed effects 3rd, 4th, and 5th models. In addition, based on the results of Kim et. al. [6] we also assumed that if there are one or more valleys here, then these are recognized in the 4th or 5th polynomial model.

IV. Discussion

Using a database of 200 different VR avatars, we were able to find evidence for the uncanny valley phenomenon of Mori et al. [1]. Contrary to expectations, however, this was not the case for the entire sample, but only when we looked more closely at different sub-categories of avatars. This means that when many different avatar categories, such as different age classes or different styles, are analyzed together, the individual graphs overlap and thus close the uncanny valley. The valley can only be created if the data is sorted precisely.

Particularly noticeable here were avatars representing children. We found that when trying to make this type of avatar more realistic, some avatars were perceived much more negatively than avatars that did not try. Avatars that received a lower human-likeness score of under 40 out of 100 points for uncanniness were significantly better than those with a higher humanlike score and over 60 out of 100 points for uncanniness.

This phenomenon cannot be replicated in the other age groups. We assume this is due to the proportions of the avatars. Because the unrealistic avatars in particular have a significantly larger head than the realistic avatars, which have normal proportions here. We were able to make this observation with the caricatures of real personalities such as former presidents of the USA. An uncanny valley can also be recognized here and, like the avatars representing children, these are mainly avatars with unusual proportions. This could be since an attempt was made here to depict real people and by increasing the similarity the avatars fall into uncanny again. However, categorization has also significantly reduced the number of data sets, which means that the coefficients of determination are very low and therefore not highly representative. This in turn can lead to the fact that certain types of avatars could not be evaluated to their full extent and an even larger database with many more avatars is needed. For example, a minimum number of avatar types could be predefined to ensure representativeness and thus fill the database more evenly. Furthermore, as a pilot study, the work focused on the development of the database and the study itself. As a result, the chapters Results, Discussion and Conclusion are somewhat shorter. The focus on the results and thus the length of the respective chapters will change in the subsequent papers.

V. CONCLUSION

With the drastic development of virtual reality and the constantly growing environment and possibilities it offers us, human-like avatars are also becoming an important topic that will affect us in the coming years. Even now, avatars from different areas are being rated according to their appearance and the term uncanny valley is being used more and more frequently. Based on the results of this study, we were able to find out that the uncanny valley is not an all-encompassing phenomenon in relation to VR avatars. Instead, the uncanny valley can only be found when taking a closer look at the individual subcategories of avatars. For example, if you take all the avatars in this database, there is an increasingly linear development between human-likeness and uncanniness, with the uncanny factor decreasing as human-likeness increases. However, if you look at certain subcategories, you can see a valley. This observation can also be observed in other categories, which leads us to assume that an overlap between the individual categories means that the uncanny valley is closed and thus balanced out by different avatars. In order to confirm this assumption, further and possibly even larger-scale studies than this one are needed. And by continuing to develop this database, we want to make this possible for everyone.

ACKNOWLEDGMENT

This work was supported by the Würzburg-Schweinfurt University of Applied Sciences. We would like to acknowledge the work of Kim et al. and the developers of the ABOT database. Their work can be found in [3] and [6].

REFERENCES

[1] M. Mori, K. MacDorman, and N. Kageki, “The Uncanny Valley [From the Field],” IEEE Robot. Automat. Mag., vol. 19, no. 2, pp. 98–100, 2012, doi: 10.1109/MRA.2012.2192811.

[2] B. Verplank, A. Sutcliffe, W. Mackay, J. Amowitz, and W. Gaver, Eds., Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques. New York, NY, USA: ACM, 2002.

[3] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, “What is Human-like?,” in Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago IL USA, 2018, pp. 105–113.

[4] M. B. Mathur and D. B. Reichling, “Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley,” Cognition, vol. 146, pp. 22–32, 2016, doi: 10.1016/j.cognition.2015.09.008.

[5] K. F. MacDorman and H. Ishiguro, “The uncanny advantage of using androids in cognitive and social science research,” IS, vol. 7, no. 3, pp. 297–337, 2006, doi: 10.1075/is.7.3.03mac.

[6] B. Kim, M. Bruce, L. Brown, E. de Visser, and E. Phillips, “A Comprehensive Approach to Validating the Uncanny Valley using the Anthropomorphic RoBOT (ABOT) Database,” in 2020 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 2020, pp. 1–6.

[7] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, “All robots are not created equal,” in Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques, London England, 2002, pp. 321–326.

[8] C. Bartneck, T. Kanda, H. Ishiguro, and N. Hagita, “Is The Uncanny Valley An Uncanny Cliff?,” in RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication, Jeju, South Korea, 2007, pp. 368–373.

[9] K. Sasaki, K. Ihaya, and Y. Yamada, “Avoidance of Novelty Contributes to the Uncanny Valley,” Frontiers in psychology, vol. 8, p. 1792, 2017, doi: 10.3389/fpsyg.2017.01792.

[10] D. Chattopadhyay and K. F. MacDorman, “Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley,” Journal of vision, vol. 16, no. 11, p. 7, 2016, doi: 10.1167/16.11.7.

[11] V. Schwind, K. Wolf, and N. Henze, “Avoiding the uncanny valley in virtual character design,” interactions, vol. 25, no. 5, pp. 45–49, 2018, doi: 10.1145/3236673.

[12] R. K. Moore, “A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena,” Scientific reports, vol. 2, p. 864, 2012, doi: 10.1038/srep00864.

[13] Oxford English Dictionary. [Online]. Available: https://www.oed.com/ (accessed: Apr. 7 2024).

[14] K. Mortensen and T. L. Hughes, “Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature,” Journal of general internal medicine, vol. 33, no. 4, pp. 533–538, 2018, doi: 10.1007/s11606-017-4246-0.

[15] D. J. Ahler, C. E. Roush, and G. Sood, “The micro-task market for lemons: data quality on Amazon's Mechanical Turk,” PSRM, pp. 1–20, 2021, doi: 10.1017/psrm.2021.57.