Can we estimate the perceived comfort of virtual human faces using visual cues?

Greice P. Dal Molin, Felipe M. Nomura, Bruna M. Dalmoro, Victor F. de A. Araújo, Soraia R. Musse
Graduate Programa in Computer Science, School of Technology, Pontifical Catholic University of Rio Grande do Sul
Porto Alegre, Brazil
Email: soraia.musse@pucrs.br

2021 IEEE 15th International Conference on Semantic Computing (ICSC) | 978-1-7281-8899-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSC50631.2021.00085

***Abstract*—The sense of strangeness (or discomfort) perceived in certain virtual characters, discussed in Uncanny Valey (UV) theory, can be a key factor in our perceptual and cognitive discrimination. Understanding how this strangeness happens is essential to avoid it in the process of modeling virtual humans. In this paper, we investigate the relationship between images features and the discomfort that human beings can perceive. We extract image features based on Hu Moments (Hum) and Histogram Oriented Gradient (Hog). The saliency detection is also extracted in the specific parts of the virtual face. Finally, a model using Support Vector Machine (SVM) to provide binary classification is suggested. The results indicate accuracy of around 80% in the image estimation process comparing with subjective classification. As a contribution, some areas may benefit from this study for avoiding the creation of characters that may cause strangeness, such as the games, conversational agents and cinema industry.**

I. Introduction

The area of Computer Graphics (CG) has stood out in the sophisticated creation of environments and characters. The similarity to the real world surprises both researchers and users of the area of entertainment and of areas such as health, laws, among others. Assessing the perceived quality of the content of images and videos is important in processing this data in various applications, such as films, games, but also platforms that use images to communicate relevant information [1]. The area of visual perception is highly complex, influenced by many factors, not fully understood and difficult to model and measure [2]. For these reasons, subjective assessments are still very used, in which a group of human viewers evaluate qualitatively the images/videos [3]. However, some problems may require quantitative assessment, when subjective analysis is not possible. Shahid et al [1] provide a model to assess the perceptual quality of images through their properties and characteristics. The perceptual problem we are interested on investigating in this paper is known as the theory of Uncanny Valley. In the 1970s, Japanese robotics professor Masahiro Mori realized that when human replicas behave very similarly, but not identical to real human beings, they provoke disgust among human observers because subtle deviations from human norms make them appear frightening. He referred to this revulsion as a drop in familiarity and the corresponding increase in strangeness as Uncanny Valley [4]. In recent decades, the Uncanny Valley has come to be considered in CG characters, whose image analysis can be inspired by the characteristics of the human visual system (HVS), as mentioned by Sanches et al [5]. Authors created a mechanism extraction of guidance resources based on physiological studies of visual perception, seeking to capture the user's subjectivity in relation to the image. Prendinger [6] defines that studying UV in the context of Computer Graphics images is a relevant case study. There is a consensus that there are characters that cause a bad feeling even though the best techniques are used, as well as characters that cause a good feeling even if advanced techniques are not used, as described in [7]. The main question of this work is to investigate if image features on the face of CG characters can help to define when images can cause strangeness.

II. Related Work

According to Hanson et al. [8], the new challenges for computer animations and simulations point to contextualizing conversations, human and environmental perceptions, as well as having control over motives or decisions. Therefore, there was a concern to evaluate the appearance and behavior of Computer Graphics (CG) characters through the Uncanny Valley, being associated with human similarity that can be used for a wide range of applications, as indicated by Tinwell et al. [9]. Through the various studies in this area of animation, some characteristics in CG characters already show greater strangeness to the human being, when evaluated, as follows: rigid or sudden movements, in the study by Bailenson et al. [10]; lack of human similarity in the speech and facial expression of a character, in the studies by Tinwell et al. [9]; lip synchronization error, according to studies by Gouskos et al. [11]. Human perceptions about characters created with CG in relation to the effect created by Uncanny Valley theory have been extensively studied in Computer Graphics. For example, in the work of MacDorman and Chattopadhyay [12], the authors tried to determine whether reducing realism in visual characteristics would increase the effect of UV. In the work of Flach et al. [7], the authors evaluated the effects of Uncanny Valley theory on human perception of CG characters from different media, such as movies and games. The authors worked with images and videos to obtain the participants' perceptual data. Results indicated comfort increases as characters are more realistic and comfort declines in videos compared to static images.

DOI 10.1109/ICSC50631.2021.00085

III. The Proposed Model

The purpose of our model is to infer whether the character cause strangeness/discomfort or not on people. Firstly, the Haar Cascade method is used to detect the faces and parts of the face, such as: eyes, eyebrows, mouth, jaw. The descriptors Hu Moments and Hog are used to generate the vector of characteristics of the entire face.The saliency function is used to show a part of the face that stands out to extract features with the descriptors. We use the PCA as reduction of dimensionality to define the most relevant variables of the vector of characteristics. Finally, the Support Vector Machine (SVM) model is used for the binary classification of the detected faces.

A. Dataset of images/videos

Our selection of characters is based on the work of Flach et al. [7], who analyzed, with subjects, the perception of comfort obtained when watching characters created with CG (films, games and computer simulations). We use images and videos of the same 10 characters, as shown in Figure 1 from (a) to (j). In addition to the characters from Flach's work, we included more recent CG characters, as shown in Figure 1 from (k) to (v). To guarantee the variation of human similarity, some of chosen characters represent a human being in a cartoonish way (q), (s) and (u), and other are more realistic, as (m), (n), (v), (r), (k) in Figure 1. To obtain human

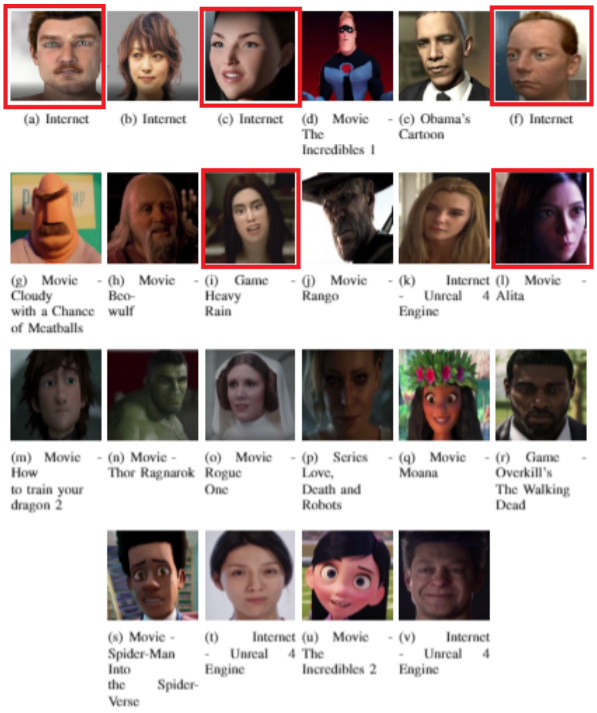

Fig. 1: All characters used in this work. From (a) to (j): characters used in Flach et al. [7]. Remaining ones are chosen for this work. The characters with a rectangular frame in red caused discomfort in the research carried out.

perceptions of comfort about the characters created with CG, we asked to participants the question "Do you feel some discomfort (strangeness) looking to this character?" which possible answers "YES" and "NO", to participants. We use Google Forms and recruited participants in social networks. All participants were asked whether they agreed to give their answers and personal information to our survey in relation to age and gender. The characters (shown in Figure 1) were randomly presented to the participants through images and short videos. We obtained 119 participants answers, 42% of which were women and 58% of men, and 77.3% being less than 31 years old and 33.7% being 31 or more years old. To compose our dataset, we firstly extracted the frames from the 22 videos (one short movie for each character illustrated in Figure 1) and removed those frames which do not contain the face of the character to be analyzed. We also removed faces from the frames that did not refer to the character in question. This process resulted in 5,799 images. Each image is classified in one of the two classes: (1) images that generate discomfort on the subjects and (0) images which do not generate. In order to classify the images in the dataset, we used simply the majority of answers YES and NO, for each image. The images are divided as follows: 2,995 images classified with major part of "YES" (class 1) and 2,804 from class 0. Concerning the characters, 5 of them generate strangeness on user perception and 16 do not, because the character (u) of the Figure 1 did not have his face recognized by the Haar cascade method.

B. Pre-Processing Data

We performed two main processes in order to prepare data to be used in our classification method, as detailed next. We implemented our method using OpenCV, scikit-learnand dlibin this processes.

*1) **(A) Face detection**:* The method used for face detection was proposed by Paul Viola and Michael Jones [13]. This method detects a face, and also the parts of the face such as mouth, middle of mouth, right and left eyes, right and left eyebrows, nose and jaw. If there is no detected face, the image is discarded, as mentioned earlier. The only character that had no face detected was from the movie Incredibles 2, as shown in (u), in Figure 1.

*2) **(B) Saliency detection**:* We computed the salience map of each of 5,799 images. This method is based on the difference between the center and the outline of the image. It is a form of extraction that has a quality that attracts the attention of human beings, as studied by Jia et al. [14] and You et al [15]. The feature descriptors presented next are generated in images with and without the extracted salience in order to assess its usefulness in detecting images that present strangeness on human perception. The method used was the Fine Grained saliency from the OpenCVlibrary with the default parameters.

C. Features Extraction

At this stage, we used Hu Moments [16] and Histogram of Oriented Gradients (HOG) [17], because they are two features widely used in the computer vision area, and were not used in the literature for the detection of perception of

strangeness, as far as we know. Hu Moments is used with its default parameters [16], implemented using OpenCV.This descriptor generates 7 moments regardless of image size. The other descriptor used was the Histogram of Oriented Gradients (HOG) [17]. We consider the detection window with gradient voting into 9 orientation bins and 64x64 pixels blocks of 1x1 pixel cells, generating the descriptor of the image characteristics to be used. Hog was implemented using scikit-learnThis descriptor generates a feature vector depending on the size of the image. For this specific dataset, we obtained a maximum of 9 characteristics per image related to the character's face. Therefore, PCAwas used to detect the most relevant variables between the characteristic vectors of HOG and Hu Moments, defining in 95% the sum of the variables' accuracy as being relevant. For model training, the two combinations of variables were tested - using PCA and not using PCA - and the results are compared.

D. Training, Testing and Validation Process

To perform the train, test and validation, we use leave-p-out cross-validation, where $p = 2$, i.e. using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set of p observations and a training set.The Support Vector Machine (SVM) model is used with three different kernels: linear, Radial-basis function (RBF) and polynomial. A tuning of the hyper parameters is also made through Grid Search when kernels RBF and polynomial are used. The SVM model was implemented using SklearnThe values used in the Grid Search parameters for the RBF kernel are for the gamma vector = $[0.3 * 0.001.0.001.3 * 0.001]$ and for the variable C = $[50., 100., 200.]$. In the case of the Polynomial kernel the gamma values are showed in the vector $[0.001, 0.01]$, the degree variable with the values $[2, 3, 4]$ and the parameter coef0 which is a kernel projection parameter, the values $[0.5, 1]$. Obtained results are stored in a .csv file, indicating if the entire face or parts of the face (as defined earlier) was used, as well as whether salience and PCA were used. The SVM model and the chosen kernel are also shown. Then, the values of precision, recall, F1 score, accuracy and time spent are also stored to further facilitate the selection of the main model.

IV. Experimental Results

It was evaluated the impact of working with the entire face or its separated parts in the training phase. Obtained results did not present significantly difference regarding accuracy. Due to this fact, it was used the entire face on evaluations showed. We number the kernels as follows: *1)* Linear pattern with data; *2)* Radial basis function (RBF) and *3)* Polynomial. The suggested model returned 24 executions, which correspond to the features (Hu moments or HOG), with the detection or not of the saliency, with or without the reduction of dimensionality through the PCA and using the three kernel functions (linear (1), RBF (2) or polynomial (3)). In addition, we use leave-p-out cross-validation, where $p = 2$, i.e. using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set of p observations and a training set, resulting in 1920 runs. As metrics we used F1-score. The F1-Score metric indicates 4 sets of implementations with F1-Scores values between 60% and 80% when classifying the characters in classes 0 (does not cause strangeness) and 1 (causes strangeness). The best implementation uses the Hu Moments feature, the polynomial kernel, does not include saliency of data and uses the dimensionality reduction, generating a F1-Score of approximately 80%. Notice that values correspond to the average of obtained values for the 80 scenarios (5x16 in cross-data test). For the two implementations that presented F1-Score 76% and 80% the time spent was 1560 seconds and 1335.08 seconds, respectively. For the other two implementations in which the F1-Score is 64% and 75%, the time is approximately 30 and 60 seconds, respectively. Considering a compromise between accuracy and computational time, we can choose the best accuracy (80%) and third best computational time (1335.08 seconds), i.e. using Hu moment, without saliency, Polinomial kernel and PCA. Yet, another possibility can be the third best accuracy (65%) with a very small computational time (60 seconds), i.e. same configuration however using Linear kernel.

Using the best accuracy implementation obtained in last section, we present results w.r.t. characters classification. Firstly, we investigate the 5 characters that cause strangeness to people (highlighted in red in Figure 1). Predictions are made using 80 runs (16 characters that generate comfort and 5 characters that do not) in the cross-validation test. Table I shows the amount of frames extracted from the videos of the 5 characters that cause strangeness in subjective evaluation. Those frames contain only the face of the 5 characters to be predicted in the implementation of this work. In a subjective evaluation, we notice that facial expressions together with the movement of the head and body can contribute to a distortion of visual characteristics, resulting in a possible strangeness to human eyes. It is possible to notice that effect in the Table I, that the character (f), that moves a lot, as in the case of characters (a) and (l), that move little. Tinwell et al [9] already make references in their work on the issue of facial expression and also on the movement of bodies, as important

TABLE I: Number of frames extracted from the videos of the 5 characters that cause strangeness in subjective evaluation and percentage of class 0 and class 1 after prediction by the implementation model. These characters correspond to the highlighted characters in Figure 1

Characters	Number of Frames	% Class 0	% Class 1
character (a)	1786	20.00	**80.00**
character (l)	45	2.36	**97.64**
character (c)	784	20.03	**79.97**
character (f)	131	44.65	**55.35**
character (i)	249	13.07	**86.93**

factors for detecting strangeness. Therefore, we also evaluated the remaining 16 characters that do not cause discomfort to people, according to subjective evaluation. Table II shows the amount of frames extracted from the videos of such 16 characters. As before, these frames contain only the face of the analyzed characters, to be classified in our work. For the 16 characters that do not cause discomfort to people, only 4 of them were incorrectly classified as belonging to class 1, as can be seen in the Table II. So, obtained error rate is 25% in the 16 characters that do not generate discomfort in subjective evaluation. So, considering the full dataset of 21 characters (and 5799 frames), we obtained accuracy of approximately 81% in characters classification.

TABLE II: Number of frames extracted from the videos of the 16 characters that not cause strangeness in subjective evaluation and percentage of class 0 and class 1 after prediction by the implementation model. These characters correspond to the highlighted characters in Figure 1

Characters	Number of Frames	% Class 0	% Class 1
character (c)	610	**97.44**	2.56
character (b)	552	**99.80**	0.20
character (v)	427	**98.75**	1.25
character (t)	402	**67.58**	32.42
character (m)	207	**75.59**	24.41
character (h)	175	**99.57**	0.43
character (o)	145	20.36	**79.64**
character (n)	80	**76.61**	23.39
character (k)	72	**99.03**	0.97
character (p)	63	20.00	**80.00**
character (s)	34	**99.72**	0.28
character (d)	21	0.06	**99.94**
character (r)	15	**99.83**	0.17
character (q)	1	20.00	**80.00**

V. Final Considerations

Starting from the conceptualization of Tumblin and Ferwerda [18] that perception is a set of processes that actively build mental representations of the world, we investigate the visual features as a way to detect strangeness perceived by the human beings when observing CG faces. We use SVM for binary classification of the characters and compared with the estimation of strangeness, as perceived by subjects. After identifying the best model, using Hu moments, without saliency, Polinomial kernel and PCA for dimensionality reduction, predictions were made for the 21 characters. The results, according to Tables I and II, present an accuracy of 80% in the predictions.

As future work, we want to increase our dataset, because it has only 21 characters, to have more data to be tested. In addition, we want to work with temporal information, which may indicate strangeness percentages on parts of the face and the body.

Acknowledgments

The authors would like to thank to Brazilian agencies CNPq e CAPES for partially funding this project.

References

[1] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on image and Video Processing*, vol. 2014, no. 1, p. 40, 2014.

[2] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, and K. M. Iftekharuddin, "A survey of perceptual image processing methods," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 811–831, 2013.

[3] S. Theodoridis and R. Chellappa, *Image and Video Compression and Multimedia*. Academic Press, 2014.

[4] M. Mori, "Bukimi no tani [the uncanny valley]," *Energy*, vol. 7, pp. 33–35, 1970.

[5] D. Sánchez, J. Chamorro-Martınez, and M. Vila, "Modelling subjectivity in visual perception of orientation for image retrieval," *Information processing & management*, vol. 39, no. 2, pp. 251–266, 2003.

[6] H. Prendinger, J. Mori, and M. Ishizuka, "Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game," *International journal of human-computer studies*, vol. 62, no. 2, pp. 231–245, 2005.

[7] L. M. Flach, R. H. de Moura, S. R. Musse, V. Dill, M. S. Pinho, and C. Lykawka, "Evaluation of the uncanny valley in cg characters," in *Proceedings of the Brazilian Symposium on Computer Games and Digital Entertainmen (SBGames)(Brasìlia)*, 2012, pp. 108–116.

[8] D. Hanson, A. Olney, S. Prilliman, E. Mathews, M. Zielke, D. Hammons, R. Fernandez, and H. Stephanou, "Upending the uncanny valley," in *AAAI*, vol. 5, 2005, pp. 1728–1729.

[9] A. Tinwell, M. Grimshaw, D. A. Nabi, and A. Williams, "Facial expression of emotion and perception of the uncanny valley in virtual characters," *Computers in Human Behavior*, vol. 27, no. 2, pp. 741–749, 2011.

[10] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich, "The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments," *Presence: Teleoperators & Virtual Environments*, vol. 14, no. 4, pp. 379–393, 2005.

[11] C. Gouskos, "The depths of the uncanny valley," *DOI= http://uk. gamespot. com/features/6153667/index. html*, 2006.

[12] K. F. MacDorman and D. Chattopadhyay, "Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not," *Cognition*, vol. 146, pp. 190–205, 2016.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[14] H. Jia, L. Zhang, and T. Wang, "Contrast and visual saliency similarity-induced index for assessing image quality," *IEEE Access*, vol. 6, pp. 65 885–65 893, 2018.

[15] J. You, A. Perkis, and M. Gabbouj, "Improving image quality assessment with modeling visual attention," in *2010 2nd European Workshop on Visual Information Processing (EUVIP)*, 2010, pp. 177–182.

[16] J. Žunić, K. Hirota, and P. L. Rosin, "A hu moment invariant as a shape circularity measure," *Pattern Recognition*, vol. 43, no. 1, pp. 47–57, 2010.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[18] J. Tumblin and J. A. Ferwerda, "Applied perception," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 20–21, 2001.