

Subjective and objective assessment of 3D textured and non-textured Cultural Heritage Artefacts

Journal:	<i>IEEE Computer Graphics and Applications</i>
Manuscript ID	CGA-2020-01-0001
Manuscript Type:	Regular
Keywords:	Computer method for Museum/arts, human perception on 3D texture,, Human Computer Interaction

SCHOLARONE™
Manuscripts

Subjective and objective assessment of 3D textured and non-textured Cultural Heritage Artefacts

D. Gillespie, K. Welham

Abstract—The core mission of museums and cultural institutions is the preservation, study and presentation of cultural heritage content. As public expectation for more open access to information and innovative digital media increases, this is being met in cultural heritage with the creation of 3D digital artefacts using methods such as non-contact laser scanning. However, many issues need to be addressed including how the visual quality of presented dataset to the public affects their perceptual experience with the artefact. The results presented in this paper demonstrate the importance of the relationship between texture and polygonal resolution and how this can affect the perceived visual experience of a visitor. It also finds that there is an acceptable cost to texture and polygonal resolution to offer the best perceptual experience with 3D digital cultural heritage artefacts.

Index Terms— Computer method for Museum/arts, human perception on 3D texture, Human Computer Interaction

I. INTRODUCTION

The use of non-contact laser scanning first originated in the automotive and aeronautical industries (Lerch, MacGillivray and Domina, 2007), where its use in reverse engineering led to its adoption in cultural heritage for 3D documentation [3]. This 3D documentation technology offers the potential for new and exciting experiences, for visitor and researchers to interact with artefacts that are too large, or that are too damaged to be displayed or handled [3]. The 3D digital datasets that are created from non-contact laser scanning consists of points in 3D space, offering a digital representation of the real world artefact. These 3D datasets can be disseminated and interacted with via galleries and websites, allowing institutions to fully communicate their 3D cultural content to their physical or virtual visitors.

However, due to the number of points within the dataset, which can be in the millions, the datasets need to undergo various operations such as compression or simplification [4] before they can be disseminated and shared. Simplification

This paper was first submitted for review on the 3rd of January 2020. The research undertaken was supported by the National Museums Liverpool, and the EPSRC.

D. Gillespie is currently with Manchester Metropolitan University, Manchester (e-mail: d.gillespie@mmu.ac.uk).

and compression is one of the ideal solutions for the dissemination and display of 3D datasets, while maintaining their integrity. Yet, these processes may inadvertently cause degradation to the overall appearance of the 3D model, and this is true as well for their 2D texture maps. These degradations can impact on the interaction and engagement users may have with the 3D datasets, therefore there is a need to evaluate the visual appearance of the rendered simplified dataset. Especially when attempting to offer the best perceptual experience to users.

There are many metrics that evaluate the visual appeal of images produced in via computer graphics. They focus predominantly on global illumination or tone mapping [4]–[7], and how they affect the overall visual appeal of the image. They do not take into account the 3D model itself. However, the literature that does focus on 3D models, are primarily concerned with the surface of the 3D model and artefacts that may occur during various processes to the mesh. Little work has been done concerning the use of a combination of 3D model, textures and lighting for the final produced image.

This paper presents a large-scale subjective study that focuses on the impact of the entire environment including the model, textures and rendering parameters using a pair wise experiment and a subjective questionnaire involving 70 participants.

This research and the subsequent results contribute the following:

- The cost of texture and polygonal resolution of a 3D digital cultural artefact to offer the best perceptual experience.
- A study that evaluates the use of no texture versus textured 3D models, which to the best of my knowledge has not been done before.
- The effectiveness of a texture when compared to the high level of detail that is captured via laser scanning.

II. RELATED WORK

With the increased use of 3D digital replicas of artefacts within cultural heritage, very little research has been done on the perceived quality of these replicas. There are algorithms, that attempt to predict the perceived visual quality of a 3D model, but they rely on a subject quality assessment with human observers. The first subjective tests used to assess the

1 quality of 3D objects, was conducted by Watson et al. [5] and
2 Rogowitz and Rushmeier [8], which has gone on to inform
3 many studies. They tested different algorithms for the
4 simplification of 3D models at different levels. They both used
5 a rating system that asked the participant to rate the object
6 using a double stimulus versus the original [5], [8].

7 Rogowitz and Rushmeier [8], conducted two experiments,
8 one asking users to rate still images of decimated 3D objects
9 and then to rate a sequence of images showing a 3D model
10 rotating. This study alone, showed how important that lighting
11 can play in the perceived quality and that it can be changed
12 depending if an object was stationary or animated [8]. Two
13 more studies that focused on the use of subjective experiments
14 to assess perceived quality for simplified models, are by
15 Rushmeier et al. [6] and Pan et al. [4], with the use of textured
16 models. These studies focused on the how texture and
17 polygonal resolution may affect our perceived visual quality
18 of the model and how effective texture can mask artefacts.
19 Rushmeier et al. [6] discovered that a substitution of polygon
20 resolution and texture resolution are object dependent. They
21 found low resolution textures can harm perceived quality of a
22 3D object regardless of polygonal resolution, where improving
23 the texture resolution improves perceived quality. While
24 Rushmeier et al. [6] focused on the use of spheres for their
25 study, Pan et al. [4] used 3D objects and textures that were
26 captured using a 3D scanner. They proposed a subjective
27 quality metric that would contribute to perceived quality of
28 both the texture and polygon resolution. The captured data
29 from the laser scanner was constructed to provide a ground
30 truth 3D object. Simplifications were applied to both the
31 captured 3D object and texture independently, to provide 3D
32 models for the subjective test. During the testing, participants
33 were asked to rate the quality of the simplified objects when
34 compared to the ground truth. The result showed the “worst”
35 object was the most simplified object. Pan et al. [4] provided
36 an insight into the relationship between polygonal and texture
37 resolution; after a point polygonal resolution no longer affects
38 perceived quality, yet texture resolution is perceived linearly
39 [8].

40 To evaluate the visual fidelity of 3D models created from a
41 watermarking algorithm, Corsini et al. [7] proposed two
42 studies. They focus on the various artefacts that may appear
43 due to different algorithms used to watermark 3D models.
44 Using the above testing method, they acquired a mean opinion
45 score (MOS), to assess the perceived quality of various
46 algorithms used to watermark each 3D model. They also
47 proposed a perceptual metric, which combines the subjective
48 MOS with a global roughness value calculated per 3D object,
49 which is then derived into simple roughness difference based
50 on the variance of geometric Laplacian [9]. The provided
51 metric was able to provide good results, predicting human
52 perceptions of distortions on watermarked 3D models. Lavoué
53 [10] also proposed a similar study by measuring the perceived
54 quality of watermarked 3D models. The participants MOS for
55 this study [9], were used to evaluate the performance of the
56 mesh structural distortion measure (MSDM) metric, which has
57 proved to be very similar to human judgement, especially in
58 complex scenes [9].

59 In the above studies, parameters that can influence a user’s
60 perceived quality of a 3D model were identified. These

parameters included the lighting, background of the model,
texture and shading, type of objects, interaction and type of
display. All of these parameters play a major part in the design
and the study of subjective quality assessment. These
parameters were taken into consideration during the design of
the experiment for both the forced comparison and subject
quality experiments.

III. EXPERIMENT DESIGN

The large-scale experiment was conducted within a gallery
space within the National Museums Liverpool, World
museums, to evaluate user’s perceptions towards textured and
non-textured models at different resolutions. The setup
allowed visitors to interact with the original artefact and a
digital representation of the cultural heritage artefact. With the
consent of the visitors, a survey was completed regarding their
experience with the digital counterpart and the real artefact.

A. Object Selection and preparation

National Museums Liverpool has an archive containing nearly
400 3D objects. The collection including sculptures, busts,
hogback stones, reliefs, archaeological finds, a tumor and a
World War 2 bomb. The study investigates artefacts that
offered a variety in surface detail, interaction styles and
materials for the visitors to interact with. From the collection,
four objects were chosen that met the criteria for the
experiment. The chosen objects and their statistics can be seen
in table 1.

B. Stimuli Preparation and Texturing

As these objects are large 3D models, an approach similar to
that of Pan et al. [4], was taken to reduce the polygon count.
Each model was simplified using the Quadratic Edge Collapse
Decimation [11], as it allows the preservation of boundaries,
normal’s and texture coordinates. Each full resolution selected
model was decimated by 10% (experimented from 100% to
10%). Three independent reviewers compared the decimated
objects until they all noticed a difference between the
decimated models. When a difference was noticed between the
stimuli by each reviewer, a level up from that decimation level
would be appointed as the high-resolution stimuli for the
study. Details on the final selected polygon resolutions are in
table 2 and table 3.

The application of colour laser scanners is becoming more
prevalent in cultural heritage, which records the surfaces
colour, yet there are many datasets recorded with surface
materials. This is the case for the objects created from the
collections at the National Museums Liverpool. A physically
based approach for the texture creation was taken. Using
photographs taken during the scanning process, the material
were created trying to be as physically realistic as possible. A
normal, diffuse, specular and ambient occlusion map were
generated at a resolution of 2048 x 2048 pixels. The results of
the texture maps and final geometry resolutions can be seen in
figure 1. It should be noted the textures will not be identical to
the original artefact but are a close substitute.

C. Pair wise Stimuli Generation

For the pair wise experiment, the high resolution object was decimated using the Quadratic Edge Collapse [4] preserving the boundary, normals and texture coordinates to a further 70%, 40% and 10% of the high resolution, creating 4 polygon resolutions. The texture of the objects was only subjected to a loss of resolution down from 2048x2048 to 1024x1024, 512x512 being saved as PNGs to avoid any compression artefacts. A resolution of 256x256 was not chosen, as it has been shown that low resolution textures can harm the perceived quality of the 3D object regardless of the polygonal resolution [6]. 3D models without textures at various polygonal resolutions were also used in the paired comparisons, to investigate if higher polygonal resolution models could be perceived as a better perceptual experience than 3D models with textures.

64 stimuli were generated for the pair wise experiment with differing polygonal and texture resolution (4 polygonal resolutions * 4 texture resolutions * 4 different objects). The details of each objects polygonal resolutions can be found in Table 1.

D. Subjective Stimuli Generation

To create a stimulus for the subjective part of the study, a 2D image metric was used to evaluate the meshes within the scene compared to a reference model and texture resolution. The 2D metric chosen was HDR-VDP2 [12] image metric for real world scenes, catering for complications and multitudes of parameters. HDR-VDP-2 is capable of measuring the visibility and quality metric, detecting differences in images across a variety of lighting conditions [12].

Using the HDR-VPD2 metric, 68 Images were captured at different polygonal and texture

TABLE I
MODELS AND THEIR DIMENSIONS AND POLYGON RESOLUTION

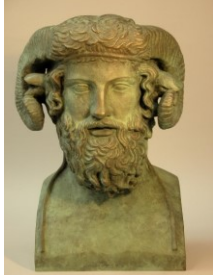



Objects	Object Image	Dimensions(mm)	Materials	Polygon Resolution	100% Resolution	70% Resolution	40% Resolution	10% Resolution	Subjective Resolution
Ammon Bust		H470 x W380 x D270	Bronze	5,040,129	1,297,076	907,952	518,830	129,706	713,392
Shakespeare Bust		H620 x W590 x D290	Terracotta	6,799,264	999,738	699,816	399,894	99,972	549,856
Egyptian Relief		H360 x W350 x D120	Limestone	498,383	136,590	174,435	99,677	24,919	136,590
Anglo-Saxon Brooch		H160 x W90 x D15	Bronze	127,5876	382,762	267,932	153,104	38,276	210,518



Fig. 1. Digital textured Representations

resolutions and compared using the code. The captured comparisons were then compared for each model, to choose a possible resolution for the subjective test. For all of the models, there was little difference between 100% geometry and 70% respectively, but there were major differences at 40% of the resolution. Thus the subjective resolution was chosen to be 55% of the original polygonal resolution. The chosen texture resolution was 1024x1024 as there was, as the difference seen in the HDR-VDP-2 metric showed was very small between the texture resolutions.

E. Setup

The study is split into two parts: one is a binary forced choice comparison experiment; and the other is a subjective experiment. The set up for the experiment would be a touch screen monitor to display the 3D models and a mouse and keyboard for input if they do not wish to use the touch screen. The interface for the experiment is minimal, showing only the 3D models within the virtual environment. The user would be able to rotate, zoom and pan the model using the provided mouse or the touch screen. To select the preferred model, the participant would select their choice on the keyboard. When a participant chooses their input, the models will automatically change to the next one, and will continue through all of the comparisons. When the comparisons end, it would change automatically to provide instructions before the subjective experiment. There are also clear and simple instructions provided at the beginning of the experiment, with a short briefing on the procedure for the experiment.

a) Pair Wise Experimental Design

Participants are asked to compare two randomly selected models, and choose either the right or the left one based on a

simple question. “Compared to this artefact, which one do you prefer?”. The semantics for this question are simply trying to reduce the bias in the results. This experiment captures data on how important the texture and polygon resolution is in relation to the perceptual experience with the 3D dataset. The users are not given a time limit on deciding between the right and left, and can freely manipulate the 3D model. The user then selects their desired choice using input from a keyboard that is provided.

To reduce the overall amounts of comparisons users would need to be made between the stimuli, a self-balancing binary tree was implemented. The self-balancing tree works off a simple assumption: If $A > B$ and $B > C$, then A is greater than C automatically, allowing for reduced comparisons. The recorded data was also screened to remove discrepancies using the ITU-R-BT.500-13 [13] protocol. Data was rejected if it was $\pm 2x$ outside the standard deviation range, or where 5% of the data was outside this range and if the values for the other values exceeded the bound of absolute difference range by 30% [13]. A small control group was used to record a full comparison matrix, resulting in 120 comparisons. Allowing for a comparison to be drawn between the full and reduced comparison tables.

b) Subjective Experimental Design

A second yet shorter experiment is also to be completed by the participant, asking the user to interact with a 3D model and with the real-life artefact. The participants were asked to answer the following questions:

- How does this 3D model and texture compare to the real object on a scale of 1 to 10? 1 being the worst and 10 being the best.
- What do you think this 3D model is made out of?
- How important is the texture for you when interacting with this 3D model?
- Would you like the option to choose to display and remove the texture from the 3D model?
- What would you prefer interacting with: the original/replication or the 3D model?
- After this experiment, would you like to learn more about the collections, or the 3D models that the National Museums have?
- Are there any additional comments you would like to make, either about the first or second part of the experiment or anything about your time here today?

The questions are designed to assess how the user perceives the quality of the decimated texture and mesh and if they understand what the material is made of. The questions regarding the use of textures, is to provide evidence of

whether or not a texture is important in the presentation of the 3D dataset.

c) *Participants*

A total of 70 participants took part in the experiment, an equal split of males and females (35M, 35F). The participants age ranged between 18 to 60, with 31 participants aged between 18 to 25, 21 being aged between 26 to 33, 11 between 34 to 41, 4 between the ages of 42 to 49 and 7 participants aged between 50 plus. Each participant had either normal or corrected vision. Participants were naïve users, visiting the World Museums Weston Discovery Centre. The users had a mixture of experiences with 3D graphics but mostly having very little experience with computer graphics. 15 Participants rated the Anglo Saxon Brooch, 15 rated the Egyptian Relief, 20 reviewed the Zeus Ammon Bust, and 20 rated the Shakespeare bust. The Brooch and relief received fewer participants, due to time constraints in the galleries and difficult nature of asking naïve users to participate in studies without an incentive. They conducted the experiments on a laptop with an Intel Core i7-2640M CPU at 2.8GHz with 8GB of RAM and a Nvidia Quadro 1000m graphics card and using a 27-inch touch screen monitor within the Weston Discovery Centre. The experiments were all conducted on different days with different models.

d) *Computing Scores*

This study used a reduced forced binary comparison test, using a self-balancing tree to reduce the number of comparisons visitors would need to make. By reducing the number of comparisons, the recorded data can be noisier than a full comparison table. However, as shown by Silva et al [14], a large number of observers can converge to be similar to the full design, while reducing the number of comparisons and time taken to complete the experiment. To calculate the scores for each model, a preference score is calculated using the formula in equation 1.

$$ps = (ta - tb) / (ta + tb) \dots (1)$$

where ta and tb are the number of times the participant preferred mesh A over B. The scores for both the reduced and full completion test would then be processed in a one-way ANOVA, to calculate a correlation between texture and polygonal resolution, and if these closely match each other. A post hoc Tukey honestly significant difference test [15] was applied to the results to show the significant results between the stimuli.

IV. RESULTS

This section discusses the results of the study, across the four 3D cultural artefacts and how their polygonal and texture resolution relationship effects human's perception of 3D digital cultural artefacts. The detailed results provide an understanding of how people rate the different resolutions, but also allows to compare the subjective score using the method from Pan et al. [4] and the HDR-VDP2 image metric compares to the full comparisons.

A. *Observers Agreement*

To scrutinise the agreement between users and their scores for the models, a Kendall's coefficient of concordance (Kendall's W) [16] was computed for each model. The produced W coefficient lies between zero and one, where zero means there is no agreement among the participants, and one there is a unanimous agreement. The results are considered significant if the p -value is extremely low ($P < 0.01$), and the null hypothesis is rejected that there is no agreement between participants. All of the P scores were below 0.01, so all of the results were significant.

B. *Paired Comparison Results*

For the forced binary comparison study, sixteen stimuli were generated per object and as can be seen in table 5, it lists the distortions that have been applied to each stimulus, and provide a guide when looking at the graphs. Each object will be discussed in turn, starting with the Brooch, Relief, Ammon and Shakespeare.

TABLE V
DETAILS ABOUT THE DISTORTIONS APPLIED TO OBJECTS

ID	Geometry Resolution	Texture Resolution
1	10% of reference resolution	None
2	40% of reference resolution	None
3	70% of reference resolution	None
4	100% of reference resolution	None
5	10% of reference resolution	512 x 512 pixels
6	40% of reference resolution	512 x 512 pixels
7	10% of reference resolution	None
8	40% of reference resolution	None
9	70% of reference resolution	None
10	100% of reference resolution	None
11	10% of reference resolution	512 x 512 pixels
12	40% of reference resolution	512 x 512 pixels
13	70% of reference resolution	512 x 512 pixels
14	100% of reference resolution	512 x 512 pixels
15	10% of reference resolution	1024 x 1024 pixels
16	40% of reference resolution	1024 x 1024 pixels

1) *Anglo Saxon Brooch*

The Anglo Saxon Brooch forced paired comparison experiment was completed with 15 museum visitors, with their data being screened using the ITU-R BT.500-13 guide [13] to remove outlier data. 4 participant's data was removed from the results analysis, due to being out of the range that was acceptable. A one-way ANOVA was used, to calculate a correlation between texture and polygonal resolution, resulting in significant results with a $p < 0.05$. A post hoc Tukey honestly significant difference criterion test was also applied to identify significant differences between the individual stimuli. The results of the post hoc Tukey honestly significant difference criterion can be seen in figure 2.

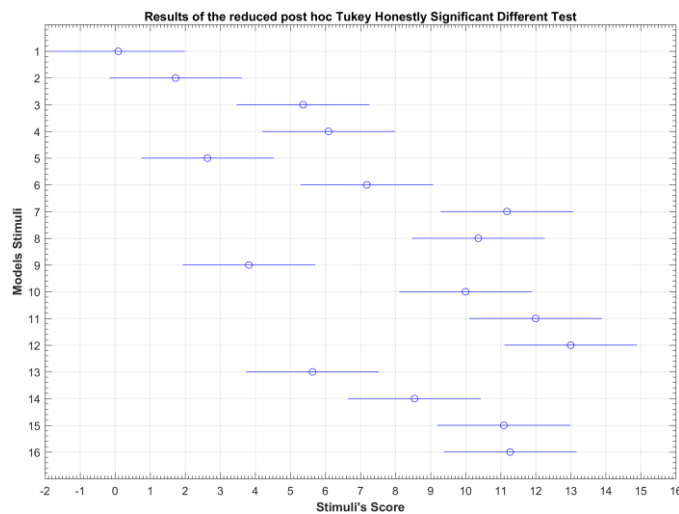


Fig. 2. Results of a post hoc Tukey Honestly Significant Different Test

At a quick glance, it appears that the lowest geometry with no texture is perceived (Tukey HSD, Score = 0.9091, $P < 0.05$) as the worst quality, with a higher polygonal and texture resolution being (ID 12) perceived as the best quality (Tukey HSD, Score = 14.89, $P < 0.05$). The increase in both texture and geometry resolution appears to affect perceived quality.

However, when looking in more detail, the belief that an increase in both texture and polygonal resolution is not so clear-cut. The lowest perceived model is the lowest geometry with no texture (Tukey HSD, Score = 0.9091, $P < 0.05$), however, it is not significantly different from the 40% stimuli with no texture or from the models with the lowest polygonal resolution and 512x 512 and 1024 x 1024 k texture resolution. While their mean scores are higher, their confidence levels overlap; there is no evidence to significantly decide which one is perceived as being of better quality. The 70% and 100% (ID 3 and 4) models (ID 3 HSD, Score = 7.257, ID 4, HSD Score = 7.984), while having mean scores less than meshes with textures, they are significantly better than models 1 and 2, yet it is not significantly different from Models (5, 6, 9, 12, 14) with texture resolutions of 512x512, 1024x1024 and 2048x 2048 with polygonal resolution of 10% and 40%. The highest rated mesh is model 12 (ID 12, HSD Score = 13) with a 100% polygonal resolution and 1024 x 1024 texture resolution, yet is not perceived as significantly different from models (7,8,10,11,15,16), which have a polygonal resolution of 70% and 100% apart from model 10 which has a polygonal resolution of 40%. There is no perceived difference between meshes with a polygonal resolution of 70% or greater with a texture applied. Though there is no significant difference, in the One Way ANOVA and Tukey HSD test, it does suggest that the increase in perceived quality is related to the polygonal resolution over texture resolution. However, there is not enough evidence to suggest that an increase in texture resolution increases the perceived quality of the 3D object.

2) Egyptian Relief

The forced comparison experiment with this artefact was completed with 15 participants, 6 users data was removed using the ITU-R BT.500.13 screening guide [13]. Their data was removed from this analysis, yet the participants still had a strong agreement among themselves. The experiments

Kendall's coefficient of concordance $W = 0.517$ and $P < 0.01$. The full comparison matrix, had a stronger agreement with $W = 0.755$ and $P < 0.01$. A One Way ANOVA was used, to calculate a correlation between texture and polygonal resolution for the Egyptian relief, resulting in significant results with a P value < 0.05 . The post hoc Tukey Honestly Significant difference test was also conducted on the data to identify significant results between the different stimuli. The results of the post hoc Tukey honestly significant difference criterion can be seen in figure 3.

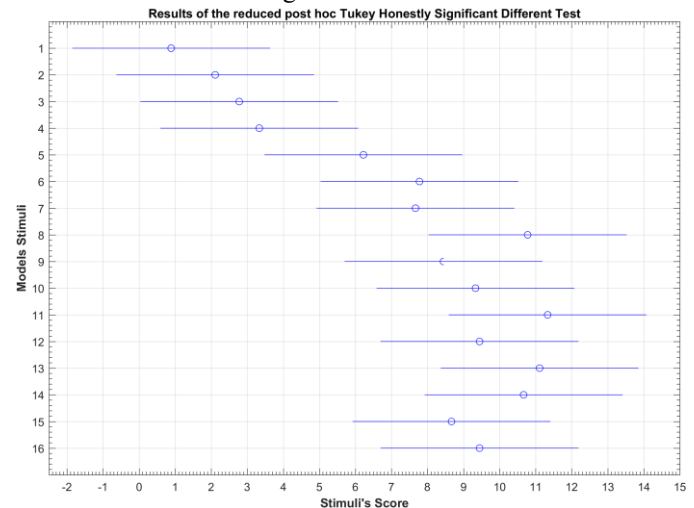


Fig. 3. Results of the post hoc Tukey Honestly Significant Different Test on the reduced data

The data gives the appearance that the worst perceived 3D model, is the lowest polygonal resolution with no texture (ID 1, Tukey HSD score = 0.8889, $P < 0.05$). The results also seem to concur with the results from Pan et al. [4], where increasing the texture resolution increases perceived quality linearly [4]. Polygonal resolution also appears to plateau, where participants cannot tell the difference between the resolutions. This is supported by the results which where the best perceived model is ID 11, which has a texture resolution of 1024x 1024px and a polygonal resolution of 70% (ID 11, Tukey HSD score = 11.3333, $P < 0.05$).

However, the results are not identical to those produced by Pan et al.[4], as it appears in the data produced for this comparison test, suggests that texture can also plateau. The lowest perceived models are the models with no textures, models ID 1,2,3,4. They have the lowest mean score, with all models being significantly perceived as better than model ID 1 and 2 apart from ID 5 (Tukey HSD, Score = 6.222) which has 10% polygonal resolution and 512x512 px texture resolution. Model 3 is similar, but there is no evidence to suggest that models with the texture resolution of 512 x512 except ID 8 (Tukey HSD, Score =10.3636, $P < 0.05$), are perceived as better. The models with textures have higher mean scores from the Tukey honestly significant test as can be seen in table 10 compared to the non-textured models, yet all of their confidence intervals overlap. There is not enough evidence to suggest that an increase in texture resolution, increases the perceived quality of a mesh. It is only possible to say that a texture improves the perceived quality for this specific mesh.

1 There are no significant differences between texture and
2 polygonal resolutions.

3 Yet, the full comparison matrix provides evidence that
4 increasing texture resolution to a point increase perceived
5 quality. The confidence levels are smaller, and show that
6 textures over 512x512 are perceived as better. Model ID 8
7 (Tukey HSD score 6.8, $p < 0.05$) is significantly better than
8 meshes 1, 2, 3 and is significantly worse than meshes 10, 12,
9 13, 14, 15, 16. However except mesh 9 (Tukey HSD score =
10 8.4444, $P < 0.05$), there is no significant difference between
11 meshes at all polygonal resolutions and 1024 x 1024 px and
12 2048 x 2048 px texture resolution. This suggests that for this
13 3D model, texture resolution plateaus similar to polygonal
14 resolution, where users cannot tell the difference in texture
15 resolution. Increasing either texture or polygonal resolution
16 after a point, will not increase the perceived quality of the 3D
17 model. However with these results, while there may be an
18 overlap between confidence intervals, there is no evidence to
19 suggest which polygonal and texture resolution is perceived as
20 the best quality or if they are perceived equally. More
21 participants would be needed for testing to increase the
22 accuracy of the results.

23 3) Zeus Ammon Bust

24 The Zeus Ammon bust, was the first of the “3D” experiences,
25 where the user had to significantly interact with the model to
26 see the full details of the 3D stimulus. 20 naïve participants
27 took part in this experiment. However, 5 participants data was
28 removed after using the ITU-R BT.500-13 screening guide
29 [13]. Unlike the “2D” interactives, there was low agreement
30 among users regarding the perceived quality of the meshes. A
31 Kendall’s $W = 0.252$ with $P < 0.01$, was computed for the
32 reduced comparison matrix. However, the P -value < 0.01 for
33 both the One Way ANOVA and post hoc Tukey HSD, showed
34 there was significant results to calculate the correlation
35 between texture and polygonal resolution. However, the low
36 Kendall’s W has led to a large amount of overlaps between the
37 individual stimuli. The lack of agreement could be due to a
38 number of reasons including; the nature of the self-balancing
39 binary tree, which can cause noisy data, especially when
40 models appear very similar.

41 However the Kendall W calculated for the full comparison
42 matrix, has a strong agreement with $W = 0.59$ with $P < 0.01$. A
43 One Way ANOVA and post hoc Tukey significant difference
44 test was conducted on both datasets. Figure 4 presents the post
45 hoc Tukey honestly significant difference test for the reduced
46 comparison, and Figure 5 presents the results of the full
47 comparison post hoc test.

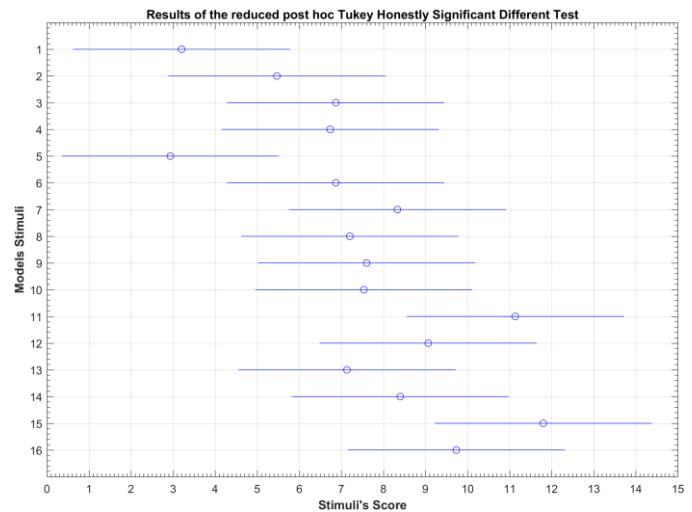


Fig. 4. Results of the Post Hoc Tukey HSD for the reduced comparisons

The results reveal that the stimuli with the worst perceived quality are stimuli 1 and 5 (ID 1 Tukey HSD score = 3.2, $P < 0.05$, ID 5 Tukey HSD score = 2.9333, $P < 0.05$). It also shows the best perceived stimuli are 11 and 15 with 70% resolution and texture resolution of 1024x 1024px and 2048x2048px (ID 11, Tukey HSD score = 11.1333, $P < 0.05$, ID 15 Tukey HSD score = 11.8, $P < 0.05$) with the smallest confidence intervals. The post hoc Tukey HSD does not reveal much information either. It reveals that even though stimuli 11 and 15 are the best perceived stimuli, they are only significantly better than stimuli 1, 2 and 5. They are otherwise are not significantly different from the other meshes, and there is no evidence to suggest they are perceived as the best quality stimuli. However, the worst stimuli is actually stimuli 5 (Tukey HSD score = 2.933, $P < 0.05$), which is significantly worse than stimuli 7, 11, 12, 14, 15, 16. There is not enough evidence to suggest that an increase in texture resolution increases the perceived quality of a stimulus. There is also not enough evidence to suggest that increasing polygonal resolution increases perceived quality either. It is only possible to say that a texture improves the perceived quality for this specific mesh. There are no significant differences between texture and polygonal resolutions apart from at the lowest and highest polygonal and texture resolution.

Yet, the full comparison matrix does provide evidence that increasing texture resolution to a point increases perceived quality. The results also provide evidence that polygonal resolution plateaus after a certain point, with scores similar between the polygonal resolutions at different levels of texture resolution. The lowest perceived stimuli is the 5th (10% polygonal resolution and 512 x 512px texture resolution) (Tukey HSD score = 1.2, $P < 0.05$), yet it is not significantly better than stimuli 1, 6,7,8,9 (ID 1 10% polygonal resolution texture), (ID 6, 7, 8 - 40, 70, 100% polygonal resolution and 512x512px texture resolution), (ID 9, 10% polygonal resolution, 1024x1024px texture resolution). The full comparison also reveals that there is no significant difference between 1024x 1024px and 2048x 2048px texture resolution. There are no significant differences, and there is not enough evidence to suggest that 2048x2048px texture resolutions are perceived as better than those of lower texture resolutions. The

full resolution also reveals that there are no significant differences between the polygonal resolution and a texture resolution i.e. 13 -14, there is no significant difference between them, and the same for 9 – 12. What was not expected was that stimuli 2, 3, 4 have high Tukey HSD scores, similar to those of stimuli with 1024x1024px texture resolutions but are not significantly different from the other stimuli except 1, 5, 6. Stimuli 3 (Tukey HSD score = 9.6) though is perceived as better than stimuli 7 and 9 as well. This trend suggests that the high resolution polygonal details captured in the mesh are either perceived quality is as good as textures and the best way to display the model. It also suggests that the textures are creating a masking effect on the 3D model, obscuring details the details of the mesh, reducing the perceived quality of the mesh. However, there is also evidence suggesting that increasing the texture resolution increases the perceived quality of the stimuli. However, due to the confidence level overlaps, between the stimuli, it is not possible to suggest what increases perceived quality of the mesh, there is not enough evidence to suggest that users perceive models with texture resolutions greater than 1024x1024px as better quality than those without textures in this experiment.

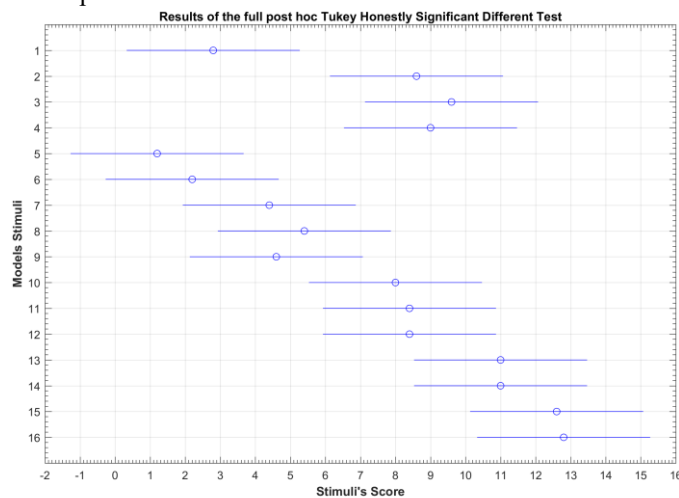


Fig. 5. Results of the full post hoc Tukey Honestly Significant Different Test

4) Shakespeare Bust

The last comparison experiment that was conducted was with the Shakespeare Bust. As with the Zeus Ammon bust, the Shakespeare experiment involved 20 naïve participants, with only 3 participants' data having to be removed following the ITU-R BT.500-13 screening guide [13]. Similar to the Zeus of Ammon experiment, there was a low agreement among participants, with a Kendall's $W = 0.344$ with $P < 0.01$ computed for the reduced comparison matrix. However, the P -value < 0.01 for both the One Way ANOVA and post hoc Tukey HSD, showed there was significant results to reject the null hypothesis. A One Way ANOVA and post hoc Tukey significant difference test was conducted on both datasets. The post hoc Tukey HSD figures are presented in figure 6 for the reduced matrix.

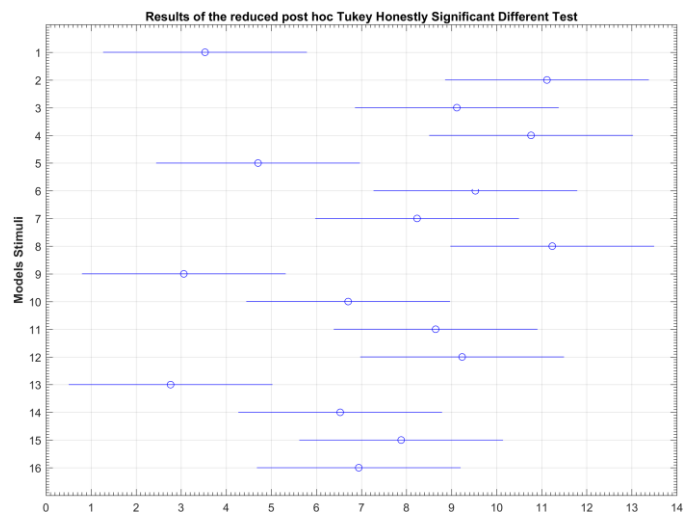


Fig. 6. Results of the reduced post hoc Tukey Honestly Significant Different Test

The data for the Shakespeare bust, like the Zeus Ammon bust, contain a lot of overlap among the stimuli shown in the reduced One Way ANOVA, though it is possible to observe a trend. The stimuli that are perceived as the worst are 1, 5, 9, 13 (ID 1 Tukey HSD score = 3.5294, $P < 0.05$, ID 5 Tukey HSD score = 4.7059, $P < 0.05$, ID 9 Tukey HSD score = 3.0588, $P < 0.05$, ID 13 Tukey HSD score = 2.7647, $P < 0.05$), which contain polygon resolution of 10% and they range across all of the texture resolutions. A trend also emerges, where the perceived quality appears linked to the geometry resolution, rather than the texture resolution. This is further supported in the full table comparison One Way ANOVA. The post hoc Tukey HSD, does reveal that there are no significant differences between the stimuli with textures, with a polygonal resolution greater than 10% except for 10 and 14 (ID 10 Tukey HSD score = 6.7059, ID 14 Tukey HSD score = 6.5294) where 10 is perceived as being worse than stimuli 2 and stimuli 14 is significantly worse than 2 and 8. It is also noted that the stimuli without textures and polygonal resolutions greater than 10% have high mean scores, but there is not enough evidence to support that they are perceived equal or better than meshes with textures. There is not enough evidence to suggest that an increase in polygonal resolution increases the perceived quality of a stimulus. It is only possible to say that meshes greater than 10% of the original mesh are perceived better than the lowest polygonal resolution.

These results are supported by the full comparison matrix, in which are very similar to the reduced comparison, except the highest resolution (ID 15), which was perceived as the best stimuli. The full matrix supports that increasing the polygonal resolution affects the perceived quality of the stimuli. In the post hoc Tukey HSD, there is no significant difference between the stimuli, where polygonal resolution is greater than 10% regardless of the texture resolution. The stimuli with no textures and polygonal resolution greater than 10%, have the highest scores (ID 3 Tukey HSD score = 11.8, $P < 0.05$, ID 4 Tukey HSD score = 11.6, $P < 0.05$), where they are in some cases being perceived as better quality than meshes with textures. However, there is not enough evidence to support that they are perceived as the best representation of the

1 cultural artefact. This does suggest as found in the Zeus
2 Ammon experiment, that it is possible texture is creating a
3 masking effect, obscuring the details on the mesh. However
4 with these results, while there may be an overlap between
5 confidence intervals, mean here is no evidence to suggest
6 which polygonal and texture resolution are perceived as the
7 best quality or if they are perceived equally.

8 C. Subjective results

9 In order to have a more complete evaluation of the 3D digital
10 representations and their perceived quality, a subjective
11 experiment comparing a digital stimulus to the real world
12 artefact was undertaken. The participants was asked multiple
13 questions which are discussed below.

14
15 1) *Question 1, How does this 3D model and texture compare*
16 *to the real object on a scale of 1 to 10? 1 being the worst and*
17 *10 being the best.*

18 This is the key question about the quality of the 3D stimulus
19 and how well it is perceived against the original artefact. The
20 user was asked to rate the stimulus from 1 to 10, on how they
21 perceived the stimulus compared to the original artefact. The
22 stimuli were rated very similarly amongst the four objects,
23 with the Anglo Saxon Brooch, being rated the highest as 7.4
24 with the lowest being the Zeus Ammon bust with 6.05. The
25 provided stimulus performed very well against the original
26 artefact, especially for heavily decimated versions of the
27 original 3D dataset. It should be noted that in some cases, this
28 stimulus for these objects was as low as 8% and as high as
29 25% of the original resolution of the original 3D dataset.

30
31 2) *Question 2, What do you think this 3D model is made out*
32 *of?*

33 Participants were asked this question to gather if the stimulus
34 was textured accurately, and if the scene in which it was
35 rendered allowed the user to accurately guess the material of
36 the object. Most participants were able to guess roughly what
37 material objects and its real world artefact was made from.
38 The Anglo Saxon Brooch had a majority of participants guess
39 the object was made from either Bronze or Gold. For the relief,
40 most participants generalised their answer to stone, with
41 a few hazarding a guess at plaster or sandstone. The same can
42 be seen in the Shakespeare bust, where participants
43 generalised they're choices to clay or stone though that is in
44 the same vein of materials that these artefacts are made from.
45 For the Zeus Ammon bust, the majority of people generalised
46 their choice to metal, though some were able to deduce that it
47 was meant to represent bronze, or contained copper due to the
48 blue patina of the texture.

49 3) *Question 3, How important is the texture for you when*
50 *interacting with this 3D model?*

51 There was a strong agreement among the 70 participants,
52 across all the objects, that the texture was quite important in
53 the interaction. Apart from two participants, that took part in
54 the Anglo Saxon Brooch, all of the other participants agree the
55 texture was either "sort of", "very important" or "quite
56 important".

4) *Question 4, Would you like the option to choose to display*
and remove the texture from the 3D model?

In conjunction with the above question, it was put to the
observer if they would like the option to see the 3D stimulus
without the texture. Similar to the above there was a strong
agreement among the 70 participants, agreeing that they would
like that option. 7 out of the 70 observers, said they would not
like that option or that they did not see it as that important.

5) *Question 5, What would you prefer interacting with: the*
original/replication or the 3D model?

Participants were asked if they would prefer to primarily
interact with; the original artefact, or the 3D digital artefact.
The answers from the study seem to suggest, that most users
would prefer to interact with both if possible.

V. DISCUSSION

This paper has presented a study, investigating how humans
perceive the quality of 3D digital datasets of real world
cultural artefacts through the use of a forced pair wise
comparison study and a subjective questionnaire. The study
has implications for cultural heritage institutions to help find
the acceptable border between polygonal and texture
resolution to offer the best perceptual experience.

The first experiment explored whether perceived quality is
linked with the texture and polygonal resolution of a 3D mesh
using differing levels of texture and polygonal resolution. The
results of this study supported studies and their claims that
texture is important to the perception of quality [4], [17], [18].
However, this study shows that an increase in texture
resolution does not increase quality linearly such as thought in
[4].

The worst perceived stimuli, was always the most extremely
decimated mesh at 10% polygonal resolution. This was
supported in the full comparison design table, which rated the
stimuli with 10% polygonal resolution consistently as the
worst with the exception of the Zeus Ammon which was the
10% stimuli with the 512x512px texture. Participants also
tended to rate models with high polygonal resolution as better
quality than those with lower. However, this is common across
other studies [4], [17], [18]. However, there was no significant
difference between meshes with polygonal resolutions greater
than 40% at any of the texture resolutions in the reduced
comparison experiment. However, the full comparison control
group, produced similar results yet there was no significant
differences between meshes with polygonal resolutions greater
than 10% with texture resolutions greater than 512x512px.

The perceived quality for each model was perceived
differently across the four objects. The Egyptian Relief, while
having an overall high polygon count is a very simple shape,
very flat with bold details. The results of its One Way
ANOVA and post hoc Tukey HSD results were similar to
what is described in Pan et al. study [4], where the texture
seems more important in the perceived quality. The score for
this object increases linearly with the worst perceived is the
10% polygonal resolution, with the score increasing linearly
before it plateaus. The material also applied to the mesh,
would also be sensitive to artefacts caused by lowering the
resolution as it can be observed more easily. This results in the

1 increased perceived quality by increasing the texture and
2 polygonal resolution.

3 This is not the case for the Anglo Saxon Brooch, which
4 seems to be the opposite. While the highest rated meshes are
5 those with the texture resolutions greater than 512x512px, the
6 scores for the meshes are very similar at their own resolutions
7 regardless of texture resolution. It appears as though both the
8 polygonal and texture resolution plateau after the 512x512px
9 texture resolution. Though as stated previously there is no
10 significant difference between meshes with a polygonal
11 resolution greater than 10% regardless of texture resolution.
12 This is the same for the Shakespeare, where scores apart from
13 the 10% polygonal resolution mesh; they all share very similar
14 scores with no significant differences between themselves.
15 This suggests that an increase in either a texture or polygonal
16 resolution increased the perceived quality of the 3D object.

17 Another observation is that non textured meshes for the busts
18 were rated quite highly, dependent on the artefact. For the
19 Zeus Ammon and Shakespeare bust, the stimuli without
20 texture stimuli scored highly against the meshes with textures.
21 This could suggest that artefacts that offer 3D interactions, the
22 use of a non-texture maybe a good alternative. The reason for
23 this trend could be due to a number of details being masked by
24 textures, unsatisfactory texturing for these models, or simply
25 participants preferred the model rendered without a texture
26 within the scene.

27 The results from the study allow us to draw conclusions on
28 the perceived quality and its relationship to polygon and
29 texture resolution. The results can be interpreted that by using
30 a method similar to that of Pan et al. [4], to appoint a new
31 lower resolution would be a visually acceptable to display to
32 the public. However, it is still possible for the model to be
33 decimated further to 40% or greater and still be acceptable to
34 the general public. The findings also suggest that a texture
35 resolution of 1024x1024px would be visually appealing
36 without the need to increase the texture resolution.

37 The second experiment focused on analysing how users
38 reacted to a created stimulus via the HDR-VDP2 image
39 metric, and their evaluation of this versus the original artefact.
40 The results from the questionnaire were similar with other
41 studies for the use of textures and their importance in the
42 perceived quality of 3D objects [4], [17]–[19]. There was a
43 near unanimous agreement between participants for each
44 object that texture was important for the display of 3D digital
45 cultural artefacts. Participants also thought there was a need to
46 have the ability to change between a textured and non-textured
47 state [19]. The results also seem to support the theory, that
48 there needs to be additional media alongside the 3D object to
49 generate interest in the object itself, and the other collections
50 within a cultural institution [19].

51 VI. CONCLUSION AND FUTURE WORK

52 The work conducted in this paper presented a methodology to
53 generate stimuli and the results of a study focusing on the
54 perceived quality of 3D cultural heritage artefacts with
55 differing levels of polygonal and texture resolution. The study
56 was under taken with 70 naïve visitors to the National
57 Museums Liverpool, where they conducted a forced
58 comparison test and a subjective study rating the perceived
59

quality of the created stimuli. The results from the
comparative and subjective experiments allowed for
interesting conclusions to be drawn regarding the perceived
quality of 3D cultural heritage artefacts.

It revealed that for each of the objects there was no significant
difference between meshes with polygonal resolutions greater
than 40% regardless of texture resolution in the reduced
comparative study. Similar results were found in the full
comparison results which revealed there was no difference
between meshes with polygonal resolutions greater than 10%
and texture resolution greater than 512x512px. This would
suggest that both polygonal [4], [17], [18], and texture
resolutions plateau. Suggesting the trend of increasing texture
and polygonal resolution may only increase perceived quality
slightly. This paper suggests that to offer a good perceptual
experience to visitors, the polygon resolution can be reduced
to 40% of the 100% resolution if following the methodology
in this paper and texture resolution does not need to be overly
large.

This paper also compared non-textured models versus
textured models, which to the best of my knowledge has not
been conducted before. Datasets that offered a 2D like
interaction style were perceived poorly compared textured
models. However, for 3D shapes that were complex in nature,
their perceived quality without a texture was rated as highly as
models with textures and in some cases perceived as the best
way to display the model.

The second experiment of this paper aimed to quantifying
how users would react to a digital replica and how it compared
to the original artefact. The results showed that the stimuli
performed well against the original artefact. It offered a good
perceptual experience for the participants, yet it was not rated
as highly as the original artefact. Participants also favoured to
either engage with the original artefact or a digital replica
instead of the digital replica by itself. This suggests that the
3D dataset still did not elicit the same ‘feelings’ as the original
artefact and the level of immersion that a monoscopic display
offers is limited.

Further work for this research, needs to be under taken
exploring if the results from the comparative and subjective
study still hold true, within a virtual environment or with 3D
printed artefacts.

REFERENCES

- [1] T. Lerch, M. MacGillivray, and T. Domina, “3D Laser Scanning: A Model of multidisciplinary research,” *J. Text. Apparel, Technol. Manag.*, vol. 5, no. 4, pp. 1–8, 2007.
- [2] F. Bernardini and H. Rushmeier, “The 3D model acquisition pipeline,” in *Computer graphics forum*, 2002, vol. 21, no. 2, pp. 149–172.
- [3] A. La Pensée, “3D Laser Scanning in 3D Documentation and digital reconstruction of cultural Heritage,” *JISC 3D Vis. Arts Network-3DVisA Bull. Issue*, vol. 4, p. 2008, 2008.
- [4] Y. Pan, I. Cheng, and A. Basu, “Quality metric for approximating subjective evaluation of 3-D objects,” *IEEE Trans. Multimed.*, vol. 7, no. 2, pp. 269–279, 2005.
- [5] B. Watson, A. Friedman, and A. McGaffey, “Measuring and predicting visual fidelity,” in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 213–220.
- [6] H. E. Rushmeier, B. E. Rogowitz, and C. Piatko, “Perceptual issues in substituting texture for geometry,” in *Human Vision and Electronic Imaging V*, 2000, vol. 3959, pp. 372–383.

- [7] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3-D mesh quality assessment," *IEEE Trans. Multimed.*, vol. 9, no. 2, pp. 247–256, 2007.
- [8] B. E. Rogowitz and H. E. Rushmeier, "Are image quality metrics adequate to evaluate the quality of geometric objects?," in *Human Vision and Electronic Imaging VI*, 2001, vol. 4299, pp. 340–348.
- [9] G. Lavoué, E. D. Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, "Perceptually driven 3D distance metrics with application to watermarking," in *Applications of Digital Image Processing XXIX*, 2006, vol. 6312, p. 63120L.
- [10] G. Lavoué, "A multiscale metric for 3D mesh visual quality assessment," in *Computer Graphics Forum*, 2011, vol. 30, no. 5, pp. 1427–1437.
- [11] M. Garland and P. S. Heckbert, "Simplifying surfaces with color and texture using quadric error metrics," in *Proceedings Visualization '98 (Cat. No. 98CB36276)*, 1998, pp. 263–269.
- [12] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, p. 40, 2011.
- [13] R. I.-R. BT, "Methodology for the subjective assessment of the quality of television pictures," 2002.
- [14] S. Silva, B. S. Santos, J. Madeira, and C. Ferreira, "Perceived quality assessment of polygonal meshes using observer studies: A new extended protocol," in *Human Vision and Electronic Imaging XIII*, 2008, vol. 6806, p. 68060D.
- [15] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (HSD) test," *Encycl. Res. Des. Thousand Oaks, CA Sage*, pp. 1–5, 2010.
- [16] M. G. Kendall and B. B. Smith, "The problem of m rankings," *Ann. Math. Stat.*, vol. 10, no. 3, pp. 275–287, 1939.
- [17] J. Guo, V. Vidal, I. Cheng, A. Basu, A. Baskurt, and G. Lavoué, "Subjective and objective visual quality assessment of textured 3D meshes," *ACM Trans. Appl. Percept.*, vol. 14, no. 2, p. 11, 2017.
- [18] J. Thorn, R. Pizarro, B. Spanlang, P. Bermell-Garcia, and M. Gonzalez-Franco, "Assessing 3D Scan Quality Through Paired-comparisons Psychophysics," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 147–151.
- [19] P. D. G. Di Franco, C. Camporesi, F. Galeazzi, and M. Kallmann, "3D printing and immersive visualization for improved perception of ancient artifacts," *Presence Teleoperators Virtual Environ.*, vol. 24, no. 3, pp. 243–264, 2015.



David Gillespie, received his EngD at Bournemouth University, while under taking a industrial placement at National Liverpool Museums. He is currently a research associate at Mnanchester Metropolitan University. His research interests include, machine and deep learning, computer vision and image processing. He can be contacted at d.gillespie@mmu.ac.uk.



Kate Welham is a Professor of Archaeological Sciences at Bournemouth University. Her primary research focus is the application of remote sensing techniques in an archaeological context. She can be contacted at kwelham@bournemouth.ac.uk