PERCEIVED QUALITY OF 3D HUMAN HEAD SCANS AT VARYING TEXTURE AND MESH RESOLUTIONS

Jan Lievens¹, Ruxandra Florea², Donny Tytgat³, Patrice Rondao Alface⁴, Peter Schelkens⁵, Adrian Munteanu⁶

¹ ETRO, Department Of Electronics and Informatics, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
² IBBT, Interdisciplinary Institute for Broadband Technology
Gaston Crommenlaan 8 b102, 9050 Ghent, Belgium
³ jan.lievens@etro.vub.ac.be
⁴ fruxandr@etro.vub.ac.be
⁵ pschelke@etro.vub.ac.be
⁶ acmuntea@etro.vub.ac.be

ABSTRACT

A 3D human head scan contains both texture and mesh data to convey the illusion of a human head. We perform subjective tests with a set of observers to distil the relative importance of texture resolution versus mesh resolution for human head scans, which is important for bandwidth limited applications such as video conferencing. The mean observer scores (MOS) obtained from the tests allow us to propose a numeric model for estimating user perception in the tested conditions.

KEY WORDS

Subjective Testing, 3D Head Scan, Texture, Mesh

1. Introduction

Assessing the perceived quality of a 3D model is a difficult task if one can only base oneself on the intrinsic properties of the texture data and mesh data that compose the model. In this paper, we investigate the perceived quality of human head scans to be used in telecommunication systems. Out of all possible 3D objects, human faces are particularly difficult to portray convincingly to human observers, due to viewers’ intricate familiarity with human faces. The overall quality of a 3D model depends on the quality of the textures and the mesh that make up the model. In communication systems, if one wants to use the available bandwidth in a system efficiently, it is then of utmost importance to have an accurate insight on how the rate spent between texture data and geometry data will affect user appreciation.

Several papers discuss the joint transmission of texture and geometry data over limited bandwidth channels. Because of the desire to transmit this possibly scalable coded data in a rate-distortion optimal manner, a function that incorporates the texture and geometry rate to represent the overall perceived quality is needed. In [1], Okuda et al. propose a progressive transmission scheme in which both texture data and mesh data are sent multiplexed in a specific order dependent on a joint metric that models the importance of a particular vertex with respect to the overall perceived quality of the textured 3D mesh. Their model for the quality metric is a weighted sum of the displaced volume by the removed vertex, and the texture similarity, both being intrinsic metrics used to quantify the distortion reduction due to the progressive representation of the data. One shortcoming is that the weights in the model are chosen freely by the user as a personal preference for either texture or geometry information.

Watson et al. were the first to show that projected domain metrics for geometry are accurate predictors for perceived quality of meshes [2]. Later, Yang et al. propose in [3] a more generic approach to joint rate-distortion optimization that is not tied to a particular transmission order or vertex decimation scheme. Their approach lies in creating a distortion surface in which the resolution of texture and mesh are both contributors to the distortion reduction. The main focus of their paper lies on finding a joint metric such that the distortion surface map may be constructed. They propose a projected 2D image-based transform domain error metric similar as Mannos-Sakrison’s filter [4], that is applied to renderings of the 3D model for the joint geometric and texture distortion. Besides being computationally intensive, they leave one future improvement path to their approach open, namely approximating the distortion reduction map with a numerical model. Similarly as in [5], Tian et al. examine joint texture/mesh coding systems and introduce a different distortion function that is a combination of both
a 2D image metric and a 3D vertex distortion metric. In their paper they show evidence of a pure 2D image approach being inaccurate as certain rendered views of a 3D model can severely mask the geometric oversimplification of a model. Their quality metric therefore incorporates both a projected geometry metric (mean squared texture deviation) and the mean squared error of the texture data. The scaling factor between the two metrics for the overall quality parameter is a content dependent constant, averaging for different 3D models around a fifty-fifty equilibrium, thus, indicating that texture and mesh data are equally important.

Finally, in [6] Pan et al. present a numerical model to predict the perceived overall quality of textured 3D meshes. Five scanned real life inanimate objects were put through a subjective evaluation. In their work, they state that their model might be improved upon by increasing the group of observers in the subjective tests and also by using a finer granularity scoring system. One of the largest subjective tests was performed on image data by Skeikh et al. in [7]. We will use this paper to set our ground rules for subjective testing.

In this work, the conditions in which we operate have a very specific set of characteristics. First, we focus on the representation of human heads specifically as being part of a system for telecommunication. Second, we also believe that the quality of human head scans is judged more critically than other more generic 3D models. Both observations have encouraged us to set up our own subjective tests built around our specific use case with the aim of obtaining a numerical model for the perceived quality of the 3D head scans.

2. Perceived Quality Of Human Head Scans

Head scans belong to the category of 3D models that are sampled directly from real world data by a semi-automatic process, as opposed to 3D models that are first conceptualized and then afterwards hand crafted by artists. The geometry of the head scans used in this paper is constructed from data acquired with stereo camera pairs, which extract 2D+z information of a scanned surface. Using multiple stereo camera pairs allows for generating a 3D point cloud representing the head. This 3D point cloud generated by the scanning algorithm forms the basis to construct the mesh-based geometry of the scanned surface. The final mesh is constructed by performing a marching cubes algorithm on the point cloud data. The use of marching cubes is particularly suited for real-time scanning applications because of its progressive scaling property. Going to finer scales (see figure 1) increases the output resolution and hence the quality of the final mesh at the expense of more computational complexity. Increasing the resolution has an impact also on the bandwidth needed for transmission of the geometry, coded or uncoded. Due to the acquisition method’s progressive resolution refinement nature, the mesh data can be coded using a scalable codec, using a lower resolution mesh to predict a higher resolution version of the mesh. This can mitigate somewhat the excessively large rate needed to send higher resolution geometry data, but is otherwise ignored further on in our
assessments of the perceived quality. Since the head scans are constructed from multiple pictures taken from different angles around the head of a real-life person, the resulting 3D model is not only composed of a very finely detailed mesh, but also overlaid with photorealistic texture data. For applications with static camera positions this texture data alone would be considered sufficient to recreate a more than adequate reconstruction of the scanned head, however in dynamic scenes where the entire environment is represented as 3D data, the quality of the geometry of the head scan is of utmost importance. The scene lighting applied in 3D environments can expose many shortcomings of inaccurate geometry representations, e.g. self-shadows generated where one would not expect them on a real human face. Furthermore, tolerance for shading artefacts depends on the overall smoothness and correctness of the face normal vectors of the mesh. Because the acquisition method described here does allow a system to generate 3D head scans with varying resolutions for both geometry data, and texture data depending on available rate, available user resources and/or available compute time, it is useful to find a function that maps each of these parameters to the final output quality.

Furthermore, as mentioned earlier, with respect to perceived quality, we believe that head scans are judged more critically than other types of models as human observers are inherently familiar with the subject to be evaluated. This is primarily why in order to establish a link between resolutions and quality, we opt to perform human observer tests as opposed to using quantitative metrics, such as the mean squared error, to determine the relative importance of texture and mesh data for head scans.

The focus of this paper is put on finding a mathematical model that aids in balancing between computational time, required network resources and perceived quality of the resulting 3D head scan. More specifically we seek to determine a function $Q = f(T_r, G_r)$, where $Q$ represents the overall perceived quality of the 3D head scan, and $T_r$ and $G_r$ denoting the resolution of the texture and geometry data, respectively. It is also important to mention that the head scans are rendered using projective texture methods. Because this rendering method does not require explicit texture coordinates to link the texture data to the geometry data, we do not consider erroneous texture coordinates as a parameter detrimental to the perceived quality, as opposed to some of the other related work.

3. Subjective Test Setup

For our subjective tests to be relevant we need a solid test set, containing multiple head scans of different people. Using the acquisition method presented above we have scanned eight different people. Seven of which were males of different ages in an age bracket ranging from twenty to forty years old and one female aged mid-twenties. While far from being an exhaustive subject set to represent the entirety of human face diversity, we believe that since we are looking purely at the effect of resolution changes this should be enough to draw conclusions on the subject of capturing 3D data. The 3D head scans for all these individuals were generated at five different resolution levels for both the geometry data and texture data, generating twenty-five combinations of head scan qualities. Regarding the mesh data, the output of the marching cubes algorithm’s scale increases exponentially, that is, the number of vertices generated increases by a factor of four with each step up in resolution. The meshes for all of the head scans at the coarsest and finest resolution level consist of approximately 5000 and 1.6 million vertices, respectively. In order to keep the scales between mesh data and texture data similar, we have decided to extrapolate the use of the exponential scale of the mesh resolution towards the texture resolution as well. Starting from the highest texture resolution level provided by our acquisition method (1600 by 1200 pixels), both the width and height of the textures are recursively downsampled by a factor two using a Lanczos-3 filter (GIMP implementation). The five discrete resolution levels were obtained by repeating this process as required for each level. The 3D head scans were captured with 4 stereo camera pairs, meaning that each 3D head scan will contain four color textures, projected on top of the generated mesh.

Because the ability to distinguish between small differences in quality of 3D objects depends on the rendered output resolution, we have decided to show the head scans rendered at the same resolution as the maximum texture resolution. If the render target resolution and the texture resolution are equal, the average rendered texel (texture pixel) area will correspond to one
screen pixel. The same goes for the geometry resolution. At the maximum available mesh resolution for our head scans, we can select a render distance from the screen so that the average area rendered by each polygon corresponds to about one pixel. We expect the evaluated quality scores given to the different head scans to saturate before the maximum quality is reached, as the differences on screen will be close to the limits of human perception. When performing the subjective testing, the reference to which the evaluation was done was the head scan composed of the highest resolution mesh and the highest resolution texture. Observers participating in our subjective tests were explained that our acquisition method provides this head scan when operating at maximum quality, and that they would be rating the other possible outputs of the acquisition method compared to this one. The test participants were not given explanations on how the quality in each of the head scans to be evaluated was going to differ from the maximum quality version. Each participant was merely put forward to give a score relative to the maximum obtainable quality, without knowing beforehand about geometry and texture resolution changes. This was deemed necessary as to not create a bias in the participants that would force them to look for small, almost unperceivable changes that would otherwise remain masked by either a higher resolution geometry or texture.

To study the effects of the resolution variations, we have decided to exhaustively sample our data set by showing each participant all combinations of texture and mesh resolution for each head scan. The participants in the subjective testing were drafted from a university environment, being members of both an engineering department in electronics and a human science department on communications. The participants had mixed familiarity with the persons depicted by the head scans, because a few of the people scanned were colleagues, but never did any participant know more than a few head scanned persons. We found that this did not influence the test results in any way.

For our practical setup we adhere to the testing standards used in video systems [8]. The tests were held in a room dedicated to the use of subjective tests, specifically equipped with an indirect light source with color temperature 6500K in order to have a fixed white balance between all conducted subjective tests. To the same effect, the test material was displayed on a monitor equipped with automatic color calibration. The participants in the subjective tests were put in front of this screen one at a time for a test session of about 30 minutes. The head scans were shown at a display resolution of 1600 by 1200 pixels, for a vertical size on screen of 24cm. The viewing distance was approximately 75cm. The observers evaluated the different head scans in pre-recorded, non-interactive 15-second test sequences. In those sequences, the frame rate was kept constant and the 3D model was shown at a fixed distance under fixed lighting conditions. As illustrated in figure 2, each head scan was displayed on screen rotating slowly around the yw-axis between 45 degrees left and 45 degrees right. The rotation started from a head-on view, going through the range of rotations left and then right, before returning to the original view at the end of the sequence. The participants in the subjective tests evaluated the test sequences in single stimulus mode. That is, before each test sequence to be scored, the reference for that head scan was shown first. The order was: show reference, show head scan under test, show reference again, show head scan under test again, cast vote.

Scoring was done on a continuous scale ranging from zero to one hundred. This range was chosen for its benefits in statistical processing as described in [9]. To help the participants familiarize themselves with the experiment (the type of data shown, the typical quality degradations, the scoring method, etc.), the participants were first exposed to a few training sequences in which the extreme quality conditions of the test range were presented. The results of these training rounds were discarded as they merely served to prepare the participants in the subjective testing to fully use the available range of scores. Even then, because the scoring range is large, not all participants could be expected to use the full extent of the range. Therefore we performed a post-processing operation on the data, by scaling the scores given by a particular user so that their lowest given score would map to zero and their highest given score to one hundred. One thing to notice is that even though the test participants were all shown the maximum quality version of a head scan as a reference many times (each scanned person’s head was tested with 25 different quality degradations), many of the participants were unable to score the reference itself the maximum score. At the highest quality points many users hesitated to give out the maximum score, even though from the results it is clear they could not differentiate clearly between the reference and that particular degraded head scan under test presented to them. Therefore in the collected test results, the average of the maximum scores will be close to the limit of the scale but not exactly one hundred.

In total 24 people completed 6 test sessions in which they performed subjective evaluation of 8 scanned heads at 5 different texture and 5 different geometry resolutions. The number of participants is well beyond the 15 needed to reach statistical significance [9].

4. Test Results And Discussion

In this section, we start of by discussing the measurements obtained from the subjective tests after our normalization procedure. These are presented above as box plots in figures 4 to 8. The horizontal axis shows the discrete geometry resolution levels and the observer scores are shown on the vertical axis as minimum, first quartile, median, third quartile and maximum obtained value in the tests.
For ease of viewing, the results are presented as one distinct box plot for each discrete texture resolution level. It is important to remark that the points on these figures correspond to a surface where the observer scores vary along two input axes, texture resolution level and geometry resolution level. This is the function we are going to model.

To continue our research we move on to using mean observer scores (MOS) of normalized data for each of the individual head scans. The results are shown in figures 9 to 16. The error bars correspond to the 95% confidence intervals of the scores at those particular mesh and texture resolution levels.

For each of the tested head scans we observe reoccurring behavior, namely a saturation part where the MOS improves only slightly by increasing the resolution of the geometry. If we denote by $G_S$ the geometry resolution from which point on the MOS saturates,
Fig. 9. MOS for head scan Maarten at various resolution levels

Fig. 10. MOS for head scan Sammy at various resolution levels

Fig. 11. MOS for head scan Bob at various resolution levels

Fig. 12. MOS for head scan Donny at various resolution levels

Fig. 13. MOS for head scan Erwin at various resolution levels

Fig. 14. MOS for head scan Freddy at various resolution levels

Fig. 15. MOS for head scan Shahid at various resolution levels

Fig. 16. MOS for head scan Marina at various resolution levels
we have the constraint for our numerical model that if \( G_r > G_s \) then \( Q(T_r, G_r) = Q_{G_S}(T_r) \) is only a function of \( T_r \).

We note that this saturation effect starts to appear as the texture resolution approaches the display resolution and the average polygon area decreases to upper single digit pixels. The reference used in our tests, the point with the highest resolution texture and geometry has detail that when rendered in our testing conditions corresponds to the pixel threshold of the screen resolution. The observed saturation effect appears distinctly before reaching this point where no automated system would be able to discern differences between the conditions under test and the reference.

The region in the graphs where the MOS scales with varying mesh resolutions shows promise to find a well matching numerical model. The model we propose for the behavior in this region is

\[
Q(T_r, G_r) = \xi(T_r)(G_s - G_r)
\]

with \( \xi(T_r) \) being the slopes of MOS with increasing texture resolution level and with

\[
\xi(T_r) = \xi(T_S) \text{ if } T_r > T_S
\]

being a constant when saturation occurs for texture resolutions greater than level \( T_S \). This captures the linear scaling of the MOS with respect to the mesh resolution level, as well as the slopes that scale with texture resolution level.

To fit our numerical model, we use the average MOS of all tested head scans (see figure 17). We combine our previous equations to express our quality metric function as

\[
Q(T_r, G_r) =
\begin{cases} 
Q_{G_S,T_S} & \text{if } G_r > G_S \text{ and } T_r > T_S \\
Q_{G_S,T_S} - \xi(T_S)(T_S - T_r) & \text{if } G_r > G_S \text{ and } T_r < T_S \\
Q_{G_S,T_S} - \xi(T_S)(T_S - T_r) - \xi(T_r)(G_S - G_r) & \text{otherwise}
\end{cases}
\]

where \( Q_{G_S,T_S} \) is the maximum MOS when both geometry and texture resolution levels are in the saturation area, and \( \xi(T_S) \) is the slope parameter of the linear fit we apply to \( Q_{G_S}(T_r) \).

Fitting our proposed model to the data of the overall average MOS of the subjective tests and keeping in mind that the observed saturation points are

\[
G_S = 3 \quad (1/16 \text{ geometry resolution}) \\
T_S = 4 \quad (1/4 \text{ texture resolution})
\]

we arrive at

\[
Q_{G_S,T_S} = 94.911 \\
\xi_{T_S} = 23.745 \\
\xi(T_r) = \xi_{T_S} - 7.1534 (T_S - T_r)
\]

as parameters for our numerical model.

To validate our proposed model, we perform an eight fold cross validation. The results are presented in table 1. For each round we include all but one person’s head scans in the training set. We train our model based on the average of those seven data points. Next, we match the MOS values predicted by our model to the one remaining data set not included in the training data.

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Pearson Correlation</th>
<th>Root Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.972</td>
<td>1.582</td>
</tr>
<tr>
<td>2</td>
<td>0.986</td>
<td>1.038</td>
</tr>
<tr>
<td>3</td>
<td>0.984</td>
<td>1.370</td>
</tr>
<tr>
<td>4</td>
<td>0.984</td>
<td>1.166</td>
</tr>
<tr>
<td>5</td>
<td>0.962</td>
<td>2.009</td>
</tr>
<tr>
<td>6</td>
<td>0.969</td>
<td>1.522</td>
</tr>
<tr>
<td>7</td>
<td>0.969</td>
<td>1.688</td>
</tr>
<tr>
<td>8</td>
<td>0.973</td>
<td>1.427</td>
</tr>
<tr>
<td>Average</td>
<td>0.975</td>
<td>1.475</td>
</tr>
</tbody>
</table>

As can be observed from the table, the average Pearson correlation coefficient is 0.975 and the average root mean squared error is 1.475, indicating that the model has good prediction properties.

We note that the model does not fully use the entire MOS scale. The maximum possible predicted MOS \( Q_{G_S,T_S} \) does not map to the maximum of the MOS range, although it represents the maximum user score obtained through human observer scores if one considers the error margin. It is important to mention that the model is not designed to predict the highest MOS point, but rather to capture the effects on the perceived quality by lowering texture and geometry resolutions starting from previously observed threshold values. In this sense, the important parameters are the slopes of the degrading MOS.

5. Conclusion

The numerical model we have obtained from our subjective tests allows us to give relative weights to the importance of texture resolution and mesh resolution in the overall perceived quality of a 3D head scan. Our model makes it possible to balance available bandwidth and desired user perceived quality. If we take into account the computational complexity increase associated with
generating a higher resolution geometry 3D head scan, the proposed model also allows making the trade-off between visual fidelity and available computational resources.

Acknowledgements

This research was carried out as part of the IBBT iCocoon project. This project is funded by the Interdisciplinary Research Institute IBBT and by the Agency for Innovation by Science and Technology (IWT-Flanders). Datasets are courtesy of Alcatel-Lucent.

References

[9] S. Winkler, On the properties of subjective ratings in video quality experiments, Proc. of International Workshop on Quality of Multimedia Experience (QoMEx), 2009, 139-144.