2014

Minimisation of video downstream bit rate for large scale immersive video conferencing by utilising the perceptual variations of quality

Pedram Pourashraf  
*University of Wollongong, pedram@uow.edu.au*

Farzad Safaei  
*University of Wollongong, farzad@uow.edu.au*

Daniel Franklin  
*University of Technology Sydney, danielf@uow.edu.au*

Publication Details

Minimisation of video downstream bit rate for large scale immersive video conferencing by utilising the perceptual variations of quality

Abstract
This paper aims at minimising the video downstream bit rate of immersive video conferencing (IVC) applications by judiciously modifying the video quality based on the relative virtual positions of participants in the virtual environment. The paper reports on the results of a user study to assess the influence of participants’ perspectives on the perceptual impact of relevant video parameters, such as resolution and frame rate. A mathematical model for video rate is proposed that expresses the total rate as the product of spatial resolution and frame rate. Results from the user study are combined with the proposed model to predict the rate parameters which will result in perceptually acceptable quality for a given user perspective. The simulation results show that by exploiting the proposed method, the downstream network load for the client can be significantly reduced with little or no impact on the perceived quality.

Keywords
variations, perceptual, utilising, conferencing, immersive, scale, quality, large, minimisation, rate, bit, downstream, video

Disciplines
Engineering | Science and Technology Studies

Publication Details
Minimisation of Video Downstream Bit Rate for Large Scale Immersive Video Conferencing by Utilising the Perceptual Variations of Quality

Pedram Pourashraf, Farzad Safaei
ICT Research Institute
University of Wollongong
Pedram@uow.edu.au, Farzad@uow.edu.au

Daniel R. Franklin
School of Computing and Communications
University of Technology, Sydney
Daniel.Franklin@uts.edu.au

Abstract
This paper aims at minimising the video downstream bit rate of immersive video conferencing (IVC) applications by judiciously modifying the video quality based on the relative virtual positions of participants in the virtual environment. The paper reports on the results of a user study to assess the influence of participants’ perspectives on the perceptual impact of relevant video parameters, such as resolution and frame rate. A mathematical model for video rate is proposed that expresses the total rate as the product of spatial resolution and frame rate. Results from the user study are combined with the proposed model to predict the rate parameters which will result in perceptually acceptable quality for a given user perspective. The simulation results show that by exploiting the proposed method, the downstream network load for the client can be significantly reduced with little or no impact on the perceived quality.

Index Terms— Video quality differentiation, video quality assessments, video conferencing, 3D immersive environments, rate model, perceptual model

1. Introduction
Video conferencing, in particular multi-party video conferencing, is now seen as an attractive alternative to face-to-face meetings. However, in traditional video conferencing systems, the network capacity grows as the square of number of participants, which will limit scalability.

An immersive video conferencing (IVC) system employs a virtual three dimensional (3D) environment and participants’ videos are displayed on the front surface of their respective avatars (Figure 1) [1]. IVC can potentially scale to a larger number of participants provided that each participant only download videos that are within their visual range. In [2], a range of techniques referred to as area of interest (AOI) management was proposed for this purpose.

In this paper, a video quality differentiation (VQD) mechanism is introduced to further improve the scalability of IVCs. The hypothesis presented in this paper is that, within the visual range of a participant, the required video quality is dependent on the relative distance and orientation of each avatar with respect to the viewer.

The contributions of this paper are as follows: (i) To verify the above hypothesis, this paper reports on the results of a subjective video quality assessment in the context of a representative 3D IVC. (ii) Informed by the results of this study, a model for adjusting video rate is developed. (iii) The impact of adjusting rate on the scalability of IVC is studied and simulation results presented.

This paper is structured as follows: in Section 2 the existing video quality assessments are studied; in Section 3 the design procedure of the subjective study and the process of subjective scores are discussed; in Section 4 the proposed models are introduced; Section 5 presents simulation results; and finally Section 6 summarises the conclusions of the research.

2. Existing Video Quality Assessments
Temporal resolution of video has been studied extensively for many years. Results show that the minimum acceptable frame rates of video is affected by many factors, including content type, viewing condition and display type. For instance, videos with fast motion require a higher frame rate to avoid “jerkiness” artefacts. However, Chen et al. concluded that the threshold of a subjective satisfaction is approximately 15 fps, although the specific value varies significantly based on the
Table 1. Spatial and temporal resolutions’ step-sizes

<table>
<thead>
<tr>
<th>Step-sizes</th>
<th>Spatial resolution(pixel)</th>
<th>Temporal resolution(fps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>352×288</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>282×230</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>226×184</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>181×147</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>145×118</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>116×94</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>93×75</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>74×60</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>59×48</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>48×38</td>
<td>1</td>
</tr>
</tbody>
</table>

The impact of spatial and amplitude resolutions was studied in [5]. It is shown that for low bit rate conditions, a low spatial resolution with smaller quantisation errors is preferred to a high spatial resolution with large quantisation error.

An assessment based on the paired comparison methodology was conducted in [7] to investigate the trade-off between the spatial resolution and temporal resolution. The outcome showed that for each fixed bit rate, when the trade-off of spatial and temporal resolution is considered, an “optimal adaptation trajectory” (OAT) can be found that ensures the maximal perceived quality. The OAT was also found to be dependent to the video content.

In [8, 9, 10, 11], three-dimensional scalability is investigated. However, in all known research in which spatial resolution is studied, the lower spatial resolution frames are always up-sampled to the larger original sizes and shown in a viewing window with fixed size. Unlike these studies, in our survey, the window (avatar size) is not fixed (see Section 4), rather its size changes based on the virtual distance and orientation of the viewer relative to the visible avatar. To the best of our knowledge, no prior works have attempted to investigate the impact of 3D characteristics of immersive environments on video quality, and none have utilised bit rate and perceptual quality models to find the minimum rate.

To obtain reference sequences, 14 ‘talking head’ videos in CIF resolution (352×288) are captured. None of the videos, apart from the last two, include an audio component. All videos are 60 seconds long and have a native frame rate of 20 frames per second. 120 target sequences are created by degrading 6 second (120 frame) slices of the 12 reference sequences (Figure 2). The degradation process affects either the spatial or temporal resolution (frame rate) of the slice. Two degradation types were included to simulate the requirements of the proposed VQD mechanism. In the degradation processes, the spatial or temporal resolution of each slice of the reference sequence is reduced by 20% relative to the previous slice. The spatial and temporal resolution step-sizes are given in Table 1.

A degradation category rating (DCR) scheme, also known as double stimulus impairment scale (DSIS), was adopted for this study [12]. However, the reference sequence and target sequence are labelled and presented next to each other inside an immersive environment. The ITU-recommended wordings were also modified to better suit an IVC environment. The five-level scale used in the study is as follows: (i) identical quality; (ii) not identical but hardly noticeable degradation; (iii) slightly noticeable degradation; (iv) noticeable degradation; and (v) unacceptable degradation.

The main goal of the user study was to ascertain and quantify the sensitivity of users to video quality parameters and obtain the minimum spatial and temporal resolutions with respect to the virtual positions and orientations of the users in the IVC such that there is no perceptible loss of visual quality.

To reduce the amount of time needed to conduct the study, the target sequence containing 10 slices of degraded videos is paired and played with the reference sequence, and subjects were free to vote while watching the videos. All videos could be viewed by each subject, which required at least 12 minutes of the subjects’ time. To maximise the accuracy and quality of the study and minimise the effects of viewer fatigue, the subjects were allowed to replay or skip any question except the last two questions.

A web-based simulation of IVC consisting of twelve questions was designed. Since it is not possible to supervise
the participants during this process, a method is developed to
detect and discard statistically invalid scores.

In each question, two avatars, a ‘reference’ avatar and a
‘target’ avatar are shown to the subject and labelled accord-
ingly (see Figure 3). The first six questions evaluate the im-
 pact of virtual distance on perceptibility of spatial and tempo-
 ral resolution degradation. In questions six to ten, the effect
of virtual orientation on the perceived video quality is studied.
The last two questions investigate the impact of the viewer’s
focal point on quality. In this paper, only the influence of vir-
tual distance on spatial and temporal resolutions is presented.

3.1. Virtual distance vs. Spatial and temporal resolution

In Question 1, reference and target avatars are located 9 me-
ters (in the virtual environment) away from the viewpoint.
The target sequence containing ten slices with ten different
spatial resolutions (100%-13.6% CIF) (Table 1) and a fixed
temporal resolution (20 fps) is applied to the front surface of
the target avatar. The corresponding reference sequence at
CIF resolution and temporal resolution of 20 fps is paired and
displayed on the front surface of the reference avatar. The
subjects are asked to indicate when they perceive any change
in the quality of the reference video with respect to the target
video. Note that the subjects were not informed of the type of
degradations that would be shown. In Questions 2 and 3, the
reference and target avatars are respectively located at 6 and
3 meters from the viewpoint, and the same process of degra-
dation is performed.

In Questions 4 to 6, a series of subjective experiments
with a fixed spatial resolution (CIF) and variable temporal res-
olutions is conducted. The frame rates of the target sequences
are dropped from 20 fps to 1 fps. As for the first three ques-
tions, the reference and target avatars were located at 9, 6
and 3 meters from the viewpoint and the corresponding reference
and target sequences were applied to the front surfaces of the
avatars. Although the sequences presented in all questions
were “talking head” videos, the video of different people with
diverse hand gestures were captured and displayed.

3.2. SUBJECTIVE SCORING METHODOLOGY

In the conducted study, the target sequence is simultaneously
presented next to the reference sequence. Hence, the scores
are relative to the reference sequence - that is, it can be as-
sumed that the reference sequence’s score is considered 5.
Therefore, if $s_{ijk}$ denotes the score submitted by subject $i$
to the slice $k$ of question $j$, then the difference scores can be
calculated as follows:

$$d_{ijk} = 5 - s_{ijk} \quad k = \{1, 2, 3, ..., 10\} \quad (1)$$

Then $z$-scores per slice are calculated:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{k=1}^{N_{ik}} d_{ijk} \quad (2)$$

$$\delta_{ik} = \sqrt{\frac{1}{N_{ik}} \sum_{k=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2} \quad (3)$$

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\delta_{ik}} \quad (4)$$

where $N_{ik}$ is the number of slices scored by subject $i$.

Then the matrix $z_{ij}$ is constructed, which corresponds to the
$z$-score assigned by subject $i$ to question $j$. A subject rejection
procedure based on the ITU-R BT 500.11 recommendation is
used [13]. According to this recommendation, the kurtosis of
the scores is calculated to detect unreliable subjects, resulting
in the removal of 20 out of 233 subjects in our study.

Figures 4 and 5 demonstrate the results of the subjective
study for the first 6 questions. As expected, regardless of the
virtual distance, recorded mean opinion scores reduce as the
spatial or temporal resolutions decreases. However, the closer
the avatar is located to the camera, the more perceptible is the
quality degradation.

As shown in Figures 4 and 5, the impact of virtual dis-
tance is more pronounced on spatial resolution than temporal.
Hence, only the impact of virtual distance is considered in
the perceptual model; nevertheless, the perception of frame
rate reduction is also affected when the avatar is located at the
furthest virtual position (Figure 5).

4. PERCEPTUAL MODEL OF THE IVC

When the 3D projection of the environment is mapped to the
2D display, avatars in the distance appear smaller than avatars
close by due to the perspective projection of the scene. Hence,
the size of the visible window showing the avatar’s video
stream is dynamic, even though the size of the user’s viewing
window is fixed. In order to find the perceptual relationship
between the virtual distance between camera and target avatar and the spatial resolution required to achieve a constant quality level, first the avatars’ projected size for different virtual distances is measured (blue circles). Using required 3D transformation matrices, a model is then obtained to predict the required spatial resolution in pixels based on the virtual distance in meters (red curve). According to the submitted subjective scores, the average spatial resolution thresholds perceived as noticeable degradation at different distances are extracted and mapped to the curve (Magenta crosses). A three-parameter exponential function is then fitted to the subjective quality scores (dashed curve).

Let $s$ represent the spatial resolution that the perceptual model predicts for an avatar at virtual distance $\beta$.

$$s = \alpha_1 + \alpha_2 e^{(-\alpha_3/\beta)} \quad (5)$$

In order to find the parameters that minimise the least square error between the vector of subjective study and the vector of fitted prediction model, the Matlab function "nlinfit" is used (Figure 6).

### 4.1. Bit-rate model

Many studies have addressed and modelled the impact of spatial, temporal and amplitude resolutions on perceptual quality [14, 15]. A mathematical perceptual model and a modelling of the bit rate in terms of the quantisation parameter, spatial resolution and frame rate is proposed in [16]. In this section, an analytical model for video bit rate in terms of spatial and temporal resolutions is proposed. The model presented in [16] is adopted and a bit rate model as a function of spatial and temporal resolutions is considered and hence the bit rate $R(s, t)$ is written as:

$$R(s, t) = R_{max}R_s(s, t_{max})R_t(t, s) \quad (6)$$

where $R_{max} = R_{max}(s_{max}, t_{max})$ is the maximum bit rate achieved by the chosen maximum spatial resolution $s_{max}$ and the chosen maximum temporal resolution $t_{max}$.

The normalised rate vs. spatial resolution (NRS) is defined as follows:

$$R_s(s, t_{max}) = \frac{R(s, t_{max})}{R(s_{max}, t_{max})} \quad (7)$$

NRS describes how the bit rate is reduced, as the spatial resolution decreases from $s_{max}$. Similarly, the normalised rate vs. temporal resolution (NRT) describes the effect of temporal resolution on the bit rate and is defined as:

$$R_t(t, s) = \frac{R(s, t)}{R(s_{max}, t_{max})} \quad (8)$$

To understand the impact of spatial and temporal resolutions on bit rate, three random talking head videos from the reference sequences of the user study are chosen. A degradation process is applied to each of the full 60 second video sequences. In the process, the spatial or temporal resolution of the reference sequences are reduced by 20% relative to the previous attempt to achieve 60 different videos with the spatial and temporal resolutions used in the study. The bit rate for each sequence is calculated. The resulting bit rates are normalised by the rate at the highest frame rate, i.e. 20 fps for that specific spatial resolution.

The curves achieved by different frame rates overlap with each other and can be characterised by a single curve. The behaviour suggests that the impact of spatial and temporal resolutions on bit rate are separable. Hence, the bit rate can be modelled as two independent functions of only $s$ and $t$.

$R_t(t)$ was shown in [17] to be a power function, explained as follows:

$$R_t(t) = \left( \frac{t}{t_{max}} \right)^b \quad b \leq 1 \quad (9)$$

Experimental data also confirmed the independence of $s$. Based on the measured data, the suggested function to model the system based on spatial resolution is:
The parameters $b$ and $d$ are obtained by minimising the mean square error between the measured and predicted rates. Since talking head videos are used in this study and the VQD system is dealing with the videos frame by frame passing the frames to the codec, the value of $b$ was approximately 1. However, according to other studies with diverse video content, $b$ has been found to vary with the intensity of motion [17]. The parameter $d$ had the value of 0.6312.

Combining Equations 9 and 10, the following overall rate model is proposed:

$$R_s(s) = \frac{s}{s_{max}}^d d \leq 1$$ (10)

$$R(s, t) = R_{max} \left( \frac{s}{s_{max}} \right)^d \left( \frac{t}{t_{max}} \right)^b$$ (11)

where $s_{max}$ and $t_{max}$ are the maximum spatial and temporal resolutions respectively and should be set based on the required application. $R_{max}$ is also the highest bit rate achieved when spatial and temporal resolutions are set to the maximum, and $b$ and $d$ are the model parameters.

To analyse the accuracy of the model, the Pearson Correlation (PC) between the measured and predicted rates are calculated. The PC of 0.9997 and 1 for prediction of the bit rate based on spatial and temporal resolution was achieved respectively, showing that the model is very accurate.

5. SIMULATIONS

In [2] a simulator was implemented to analyse the amount of bandwidth saved after applying the proposed AOI algorithms in different scenarios. In this section, the simulator is utilised to evaluate the proposed VQD mechanism. A model is obtained by combining the Equations 11 and 5 to calculate the required spatial resolution based on virtual distance. By exploiting the model and classifying the avatars into 3 spatial regions based on their virtual distances to the local client, and assigning different temporal resolutions to each region, the VQD mechanism was integrated to the simulator. The IVC in all experiments is a fixed size (100 m $\times$ 100 m) virtual environment and the simulations are performed with 100 iterations in which the clients are distributed uniformly with random orientations, unless otherwise specified.

In order to simulate a realistic scenario, translational and angular velocities are introduced to the system. The avatars are clustered around centre points, and number of centre points varies from 1 to 6. The maximum possible translational and angular velocities are 15 m/s and 180 deg/s respectively, and each client is assigned a random velocity between zero and the maximum velocity. A total of 25 avatars is clustered around different numbers of centre points, facing their respective centre points with an offset of $\pm 15$ degrees. The local client is placed as one of these centre points.

As demonstrated in Figure 9, occlusion culling is highly effective due to the high density of clients around the local client. Nevertheless, VQD can still further improve the bandwidth saving. For the case where the clients are clustered around 6 centre points, VQD achieves further savings of 43.80% compared to using AOI methods alone.

A worst case scenario for the AOI management system would be the one in which all avatars are in the visual range and view frustum of the local client. Such a scenario could occur in a virtual lecture theatre environment, where all clients are arranged on a pitched floor and hence, from the lecturer’s perspective, they are not occluded by each other.

In this experiment, the number of clients in the lecture theatre is increased from 5 to 60. The clients are positioned randomly in a virtual lecture theatre as explained earlier. As expected, the AOI methods are entirely ineffective. However, after exploiting the VQD strategy, a significant average bandwidth saving of 74.60% is achieved. This result is shown in Figure 10.

6. CONCLUSION

In this paper, a subjective study was presented, which aimed to evaluate the impact of virtual distance and orientation on perceptual quality of video. The study included 120 video sequences derived from 12 reference sequences and assessed by 233 subjects. The results showed that subjects can tolerate higher spatial degradation when the avatars are located further away from the viewpoint in the 3D virtual environment.
Based on these survey results, a perceptual model and bit rate model were proposed, which fits the measured rates very accurately, with an average Pearson correlation of 0.9998.

By combining these results, the required spatial resolution based on the virtual distance can be predicted so as to have a negligible perceptual impact on the viewer, while significant bandwidth saving can be achieved.

7. REFERENCES


