A Study of User Perception of the Quality of Video Content Rendered Inside a 3D Virtual Environment

Pedram Pourashraf, *Member, IEEE,* Farzad Safaei, *Senior Member, IEEE,* Daniel R. Franklin, *Member, IEEE*

***Abstract*—This paper reports on the results of a user study which assessed the perceptual impact of relevant video parameters, such as resolution and frame rate, based on the perspectives of participants in the virtual environment. A mathematical model for video rate is presented that expresses the total rate as the product of separate functions of spatial and temporal resolutions. Results from the user study are combined with the model to predict the rate parameters which will result in perceptually acceptable quality using the 3D features of the virtual environment. The results show that by exploiting the insensitivity of users to controlled quality degradation, the downstream network load for the client can be significantly reduced with little or no perceptual impact on the clients.**

***Index Terms*—Video quality differentiation, subjective video quality assessments, video conferencing, 3D immersive environments, rate model, adaptive video content**

I. Introduction

The number and variety of applications in which a virtual environment is combined with elements of reality are on the rise. This paper examines one such *mixed reality* scenario in which videos of people or scenes are displayed inside a 3D virtual space. This augmentation of the virtual with the real is often referred to as *augmented virtuality*, and is potentially useful in many situations, such as remote education, collaborative work, health and safety, and military training.

As an example, in an *immersive video conferencing* (IVC) system, the meeting takes place inside a virtual 3D environment, and the videos of participants are displayed on the front faces of their respective *avatars*. Avatars are free to move around in the virtual environment (typically subject to somewhat realistic physics), emulating a real-life human gathering. Within the 3D virtual space, it is also possible to show presentations, images, videos or other types of content on display boards or spaces based on the needs of the meeting. Figure 1 shows a screenshot of the iSee virtual environment, which is an example of such a system developed by our research team [1].

The IVC model provides a unified and natural context for communication that is not possible with a conventional 2D conferencing system. Moreover, due to the varying spatial relationships between users, all video streams need not be transmitted at the same bit rate to a given participant to achieve a consistent quality of experience (QoE). An IVC system can therefore scale to a potentially much larger number of participants. In the authors' prior work, a range of techniques was proposed to decrease the required bandwidth and hence increase the number of potential concurrent participants in the IVC environment [2], [3], [4].

In this paper, which is a significant extension of [3], the focus is on assessing the perceived quality of video and its relationship to bit rate, when the video is 'consumed' inside a 3D virtual space. As discussed in Section II, significant research has already been conducted on the impact of spatial and temporal resolutions on the perceived video quality. However, to our knowledge, all of these studies have considered the conventional model of video consumption, i.e., when the video is rendered as a rectangular window on a two-dimensional screen. In contrast, showing a video inside a virtual space involves the following steps: firstly, the video is applied onto the surface of an object (e.g., a face of an avatar or display board) as a *texture*. This surface then undergoes the appropriate transformations by the 3D engine to be rendered on the screen. The mapping between texture elements (texels) and the screen pixels is non-trivial and is a function both of time and the spatial domain of the video.

For example, if the object surface is being viewed from a large (virtual) distance, several texels may be mapped onto a single screen pixel. This is referred to as *texture minification* and, in essence, is a form of video down-sampling. It is reasonable to suppose that if the spatial resolution of video is reduced judiciously to match the degree of texture minification, users will not perceive any degradation in video quality; consequently, the relationship between video rate (spatial and temporal resolutions) and perceived quality should be strongly affected by the virtual distance and orientation of texture surfaces and ultimately the viewpoint of the local client with respect to other videos. By assessing and quantifying this relationship, it may be possible to adjust the video rate for each client without any perceivable effect on quality, enabling significant reductions in total video bandwidth requirements. We refer to this phenomenon as virtual quality degradation (VQD) in this paper, meaning that video quality can be degraded based on virtual distance and orientation without the effect being noticeable.

To find the perceptual threshold of the participant for detecting degradation in video quality, this paper presents the results of a subjective video quality assessment in the context of a representative 3D IVC. The study included 233 participants and 12 different questions analysing the impact of the avatars' virtual positions and orientations on the perception of spatial and temporal degradation of video quality.

The specific contributions of this paper are as follows:

1) Reporting the result of a large-scale subjective video quality assessment study that is focused on the relation-

P. Pourashraf and F. Safaei are with the Information and Communication Technology Research Institute, University of Wollongong, Wollongong, NSW 2522 Australia.

D.R. Franklin is with the School of Computing and Communications, University of Technology Sydney, NSW 2007 Australia.

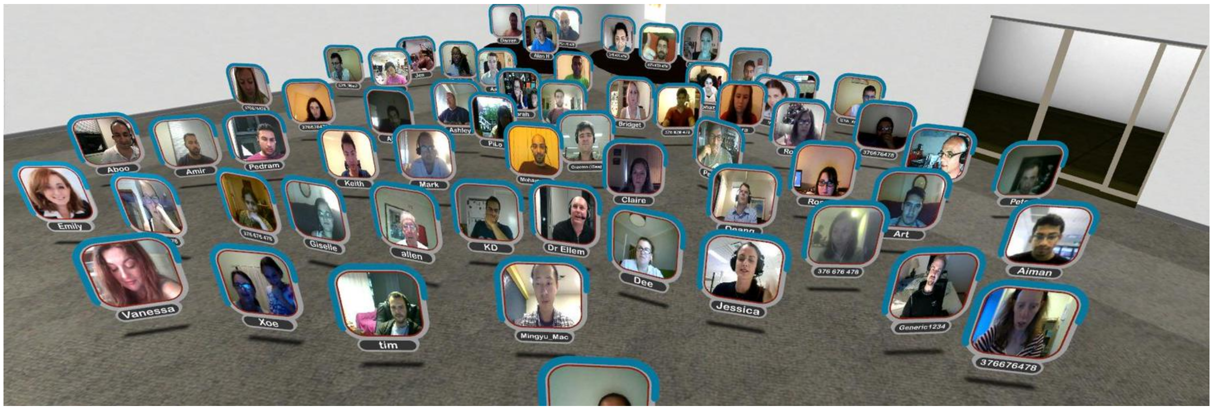


Fig. 1. Screenshot of the iSee augmented virtuality application [1].

ship between perceived quality and virtual perspective of video;

2) Developing a model to predict the required video quality and rate based on virtual distance and orientation for mixed reality scenarios; and
3) A case study in which this model is used to assess the bit rate savings that can be achieved for an IVC application by judiciously degrading video in response to variations in perspective of clients.

For the second and third items listed above, this paper assumes a *conventional* video codec, in which rate adjustment results in a more or less uniform change in quality across the spatial extent of the video. We have also developed a different rate adjustment technique, referred to as *perceptual pruning*, that can control the quality of video at a scale of arbitrary sized video blocks. This will enable more fine-grained control over video quality, especially when the video surface is viewed at an angle, resulting in greater savings in video bandwidth requirements compared to the results reported in this paper[1]

This paper is structured as follows: Section II discusses related work in the field of video quality assessment; in Section III, the design of the subjective study is described; in Section IV the subjective scoring system and the proposed models are introduced; Section V presents simulation results and analysis; and finally Section VI summarises the conclusions of the research.

II. Related work

Considerable research has been conducted into the relationship between the various factors determining the bit rate of a video stream and human perception of *quality of experience* (QoE). The bit rate of an uncompressed video stream can be controlled by changing parameters such as temporal resolution, spatial resolution and pixel amplitude sampling resolution (which determines the signal-to-noise ratio (SNR) and is controlled by the quantisation step-size (QS) or quantisation parameters (QP)). The bit rate of the compressed bit stream is usually several orders of magnitude lower due to the considerable amount of temporal and spatial redundancy in a video signal, and the preferential removal of information in the video signal which has the least impact on perception of video quality (lossy compression). A number of significant studies into the effects of varying these parameters on perceived QoE are discussed in this section.

The effects of varying temporal resolution on QoE have been extensively studied. It has been shown that the minimum acceptable frame rates of video is determined by many factors, including content type, viewing condition and display type. For instance, a video sequence containing fast motion requires a higher frame rate to avoid perceived “jerkiness” artifacts. However, for most video content, the subjective threshold for viewer satisfaction is generally regarded as approximately 15 fps, although the specific value varies significantly based on the aforementioned factors [5], [6].

In much early literature on multimedia QoE, it is widely assumed that a high frame *rate* is perceptually preferred to a high frame *quality* for content with fast motion, while for slow motion content, a reduction in frame rate has a minor perceptual impact on subjects. Several studies have demonstrated that this assumption is not entirely valid, and that the preference for frame rate versus quality is dependent on bandwidth constraints applied to the video stream, with a high frame rate preferred over high frame quality when bandwidth is limited, and the preference transitioning to frame quality over frame rate when more bandwidth is available, even for fast-motion video [7], [8].

The impact of spatial and amplitude resolution was also studied in [9]. In this study, the reference image has a spatial resolution of 320×192 pixels, and the frame rate was fixed at 30 Hz. Three different spatial resolutions (50%, 75% and 100% of the original resolution) combined with five QP values in H.263+ were analysed. The results show 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 paired comparison methodology was conducted in [10] to investigate the trade-off between spatial resolution and temporal resolution. This study utilised MPEG-4 encoding and the spatial and temporal resolution of video stimuli was varied from 40% to 100% of QCIF

[1] Note to reviewers: The perceptual pruning manuscript is currently under review by the IEEE Transactions on Multimedia and cannot be referenced. Hence, a copy is provided as supplementary material for this submission.

resolution at between 5 and 25 frames per second. The results showed that for each fixed bit rate, when the trade-off of spatial versus 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 [11], [12], [13], [14], [15], three-dimensional scalability is investigated. In all known research in which spatial resolution is studied, lower spatial resolution frames are always up-sampled to the larger original sizes and shown in a viewing window with fixed dimensions. However, the findings from these studies cannot be directly applied to the IVC environment, since the window (avatar size) is not fixed (see Section IV-B); its size changes based on the virtual distance and orientation of the viewer relative to the visible avatar. In this paper, we focus on the impact of spatial and/or temporal resolution with respect to virtual distance and orientation as described in Section IV-C. To the best of our knowledge, no prior works have investigated the impact of the unique three dimensional characteristics of immersive environments on perceived video quality. Furthermore, no previous study has considered video bandwidth constraints together with perceptual quality models to find the optimal combination of spatial and temporal resolution which minimise the required video bandwidth to meet a given perceptual quality level. This is the key objective of this paper.

III. Subjective Study Design

The subjective video quality assessment system was implemented in four parts:

1) A server-side back-end was developed in C#, which was responsible for the control and logic of the assessment system;
2) On the client side, Adobe Flash was utilised to display the recorded immersive environment, including the avatars with the reference and target video streams on their front surfaces;
3) The user interface of the system was implemented in MVC .Net, and provided a consistent interface on each of the different supported platforms and presented the progress bar, the ITU-R ACR scale, streamed the Adobe Flash files and handled other required interfaces;
4) A SQL server was used to store the scores collected from the subjects.

Since the study was web-based, full control over of the viewing environment and display configuration was impossible. However, the recorded immersive environment was configured to use the same resolution (in pixels) on all platforms and displays. The avatars with the video streams were also positioned at specific distances in each question for all subjects (described in detail in Section III-G).

A. Subject recruitment

After developing the assessment system, it was extensively tested on multiple client platforms with a variety of displays to ensure that the presentation of the user study is consistent and independent of the client platform. Once this was completed, the VQD system was deployed to a publicly accessible server. Then, electronic participation requests were sent to the students and staff of the University of Wollongong (UOW) as well as a number of other research groups in Australia. Additionally, paper posters were printed and distributed around the University in order to recruit more test subjects. A prize of Apple iPad 2 was used as an incentive to help motivate survey participation.

In the content of the request, the receivers were informed about what kinds of information will be collected from them, the means by which confidentiality of that information will be protected, how the information collected will be used in our research and a brief description of the flow of the user study. The survey was available for 40 days after the recruitment process was started and all the submitted scores in this period was stored in the database.

B. Human subjects training and testing

In order to avoid any bias, the subjects were only briefed about the goal of the experiment and the procedure of the study by a short description available on the first page of the web-based system. Additionally, a short instruction was provided above each question and the subject could read the instruction before starting the question. In the last two questions, in addition to the written instructions, an audio component was included in the video of one of the avatars, which was used to shift the attention of the subject as well as providing instructions for the question.

C. Flow of the study

In addition to the brief description of the user study available on the first page of the system, a button labelled "Go to survey" was presented, which provided access to the survey. By clicking the button, subjects were redirected to the registration page in which the identity of the subjects was verified. A range of information, including the subject's email address, gender, age group and level of education was collected. This page was equipped with a captcha protection to prevent Internet bots from registering in the survey. After a successful registration, the subject could access the first question of the study. In this stage, all questions were preloaded to RAM on the subject's machine and then released for play-back in the corresponding question, in order to prevent the videos stuttering due to network congestion.

As explained in Section III-G, each subject was free to skip or repeat any question other than the last two. If a question was skipped, no score was recorded for that particular question. In the case of a subject repeating a question, all submitted scores were stored in the database, and then in the post-processing phase the redundant answers were filtered out. In the filtering mechanism, the submitted score which was closest to the average of all submitted scores on the perceptual scale was retained and the other scores submitted by the same subject to that particular question were discarded. The system also allowed participants to leave the survey at any point and resume later.

The two final questions, 11 and 12 analysed the impact of focal point on perceptibility of spatial degradation of video quality. For this reason, the user was prohibited from repeating these questions; this ensured that the subjects could focus on the avatar, which would not be in their centre of attention area in the first attempt.

After answering the last question, the subject was informed that an email was sent to his/her registered email address in order to verify that the subject owns the email address. A link was provided in the content of the email to validate the address. Only validated addresses were used and are presented in the study.

D. Subjects' demographics

In this section, subject demographics (based on the data collected from the registration page) are analysed.

1) Distribution of subject by gender: The gender of the participants were recorded in the process of user registration. The results show that a total of 118 females (51%) and 115 (49%) males participated in our user study, with no participants registering as 'other'.

2) Distribution of subject by age: In this section, the age distribution of subjects is analysed. In the registration page, a subject could choose one of nine age bands The choices were as follows:

- Under 15
- 16-20
- 21-25
- 26-29
- 30-39
- 40-49
- 50-59
- 60-69
- 70+

The outcome revealed that the majority of subjects were from the age group of '30-39' (63 subjects) and no subject was found from the first age group which was 'under 15'. The distribution of subjects amongst different age group is presented in Fig.2.

3) Distribution of subject by education: Since the request for participation was primarily sent to Universities and research centres, 184 out 233 subjects participated in the study had education level of Bachelor's degree or higher. The distribution of subjects according to their education level is demonstrated in Fig.3.

E. Reference sequences

14 'talking head' videos were captured in CIF resolution (352×288) for use as reference video sequences. 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.

F. Target sequences

120 target sequences were prepared by degrading 6 second (120 frame) slices of the 12 reference sequences (Fig.4). 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 and network losses. 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. Let $Res_{(j)(k)}$ denote the spatial or temporal resolution of slice k of video j. Then:

$$Res_{(j)(k)} = Res_{(j)(k-1)} - \lfloor 0.2 Res_{(j)(k-1)} \rfloor \qquad (1)$$

The spatial and temporal resolution step-sizes are given in Table I.

Random frame losses that are caused by poor network condition may have significant perceptual impact due to the prediction chain in a video stream. In a typical video sequence, an single I frame is followed by a number of P frames. In other words, a single frame drop by network may translate to a burst of frame losses until the next I frame is received. Our intention

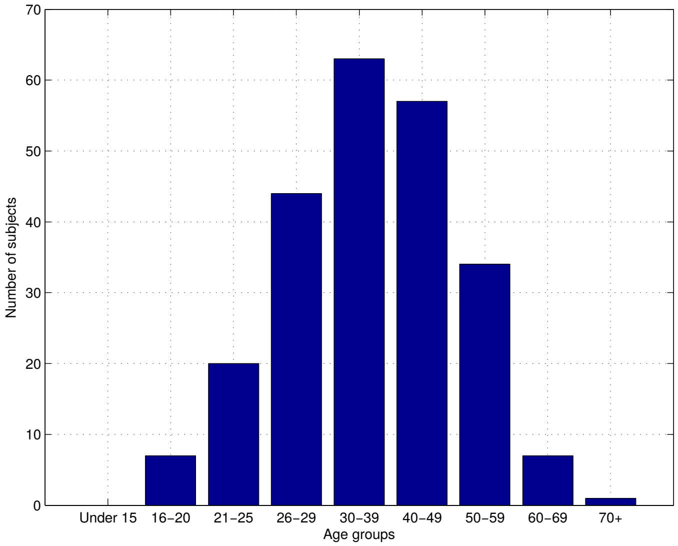


Fig. 2. Distribution of subjects by age

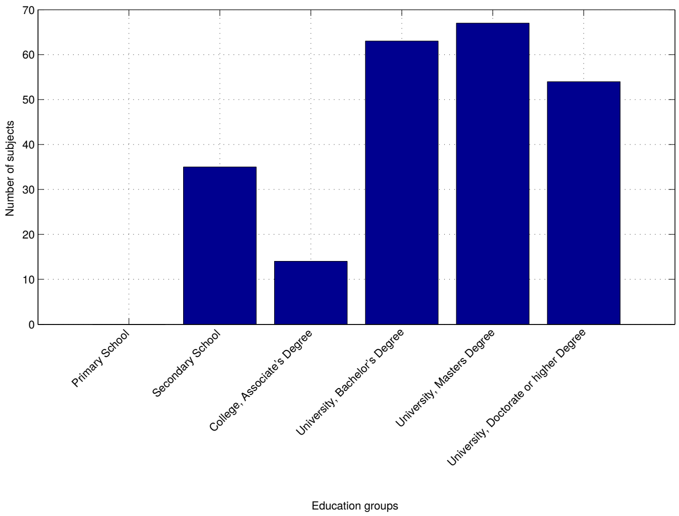


Fig. 3. Distribution of subjects by education

TABLE I
DEGRADATION LEVELS FOR SPATIAL OR TEMPORAL RESOLUTIONS USED IN THIS STUDY

Step-sizes	Spatial resolution (pixels)	Temporal resolution (fps)
1	352×288	20
2	282×230	16
3	226×184	13
4	181×147	10
5	145×118	8
6	116×94	6
7	93×75	4
8	74×60	3
9	59×48	2
10	48×38	1

is to assess the impact of reduced temporal resolution, when the frame rate reduction is intentionally introduced by the sender. In this case, a more intelligent frame reduction strategy can be adopted to avoid the above issue. To simulate this, in this research the required frame rate is achieved by dropping random frames to study the worst case scenario. However, in the real system a smart method is used to judiciously discard the right frames to obtain the best perceptual quality.

G. Subjective testing design

A degradation category rating (DCR) scheme, also known as double stimulus impairment scale (DSIS), was adopted for this study [16]. Uniquely, in this study, 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:

- Identical quality;
- Not identical but hardly noticeable degradation;
- Slightly noticeable degradation;
- Noticeable degradation; and
- Unacceptable degradation.

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.

H. Subjective testing display

A web-based user interface was developed using Microsoft MVC .Net and C#. To prevent any unintended additional distortion, the reference and target sequences for each question were transmitted frame by frame to the IVC. The immersive environment was configured to use a full-screen resolution of 790×410 pixels, and the avatars with the video streams displayed on their front surface were positioned at different virtual distances and orientations as described in Section III-I. The IVC was recorded at a rate of 25 frames per second to avoid frame dropping. To make the web-based study cross-platform and accessible to majority of the users, the recorded video streams were exported to separate Adobe Flash files for each question. To guarantee perfect playback of each segment and avoid latency due to low-speed connections, all questions were preloaded to RAM on the subjects' computers first before being released for playback. The remaining screen area around the video was white, with a progress bar and the ITU-R ACR scale (with modified wordings) displayed on the bottom. The left end of the scale was labelled "Identical quality" and the right end was labelled "Unacceptable degradation". Three equally spaced labels between these were shown: "Not identical but hardly noticeable degradation"; "Slightly noticeable degradation"; and "Noticeable degradation". A screenshot from the user study is shown in Fig.5. The subjects were asked to watch the videos, and at the point at which they perceive a change in the quality of the target avatar relative to the reference avatar, press the corresponding button. In each question, multiple buttons could be pressed and hence multiple scores could be submitted. The subjects were allowed to take as much time as needed to press the graphical representation of the "next/previous" button to navigate between questions.

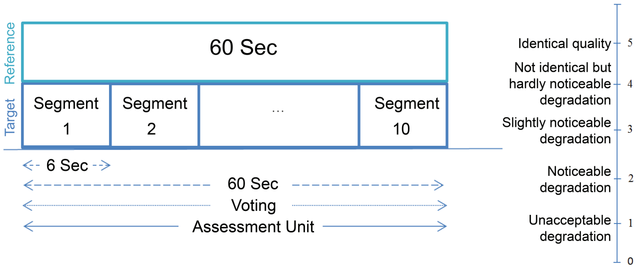
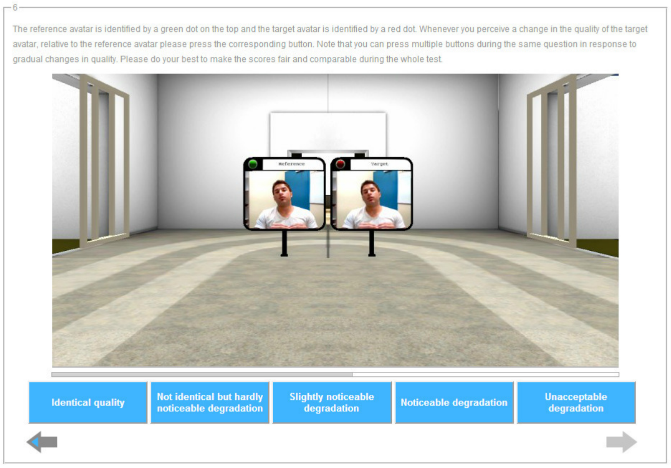


Fig. 4. Subjective assessment setup


6

The reference avatar is identified by a green dot on the top and the target avatar is identified by a red dot. Whenever you perceive a change in the quality of the target avatar, relative to the reference avatar please press the corresponding button. Note that you can press multiple buttons during the same question in response to gradual changes in quality. Please do your best to make the scores fair and comparable during the whole test.

Reference
Target
Identical quality
Not identical but hardly noticeable degradation
Slightly noticeable degradation
Noticeable degradation
Unacceptable degradation

Fig. 5. Screenshot from the user study

I. Subjective study methodology

In this research, an IVC is a 3D virtual environment allowing a large number of users to simultaneously interact and communicate using live multi-way video and audio, with their video stream displayed on the front surface of their avatar. The major bottleneck for such a system is the network transmission capacity needed to support the service. In our prior work, a range of real-time strategies were proposed to minimise the bandwidth consumption and hence improve the system's scalability [2]. The main goal of this study is to extend these results by finding the minimum spatial and temporal resolution with respect to the virtual position and orientation of the users in the IVC such that there is no perceptible loss of visual quality.

In order to determine these minimum resolutions, a subjective user test consisting of twelve questions has been designed. In each survey question, two avatars are shown to the subject. The reference avatar is indicated by a green dot on the top and the target avatar is indicated by a red dot. Both avatars are also labelled accordingly. In the first six questions, the impact of virtual distance on perceptibility of spatial and temporal resolution degradation is analysed. From question 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 changes in perceived spatial quality.

1) Virtual distance vs. spatial resolution: In Question 1, reference and target avatars are located 9 meters away from the viewpoint in the virtual environment. The target sequence contains ten slices, each at a different spatial resolution (100%-13.6% CIF) (Table I) and a fixed temporal resolution (20 fps), and is displayed on the front surface of the target avatar. The corresponding reference sequence at fixed CIF resolution and fixed temporal resolution of 20 Hz is displayed on the front surface of the reference avatar. The subjects are asked to press the appropriate button at the point where they perceive any change in the quality of the reference video with respect to the target video. Subjects were not informed of the type of degradations that they would be shown.

In Questions 2 and 3, the reference and target avatars are respectively located at virtual distances of 6 and 3 meters from the viewpoint, and the same process of degradation is performed.

2) Virtual distance vs. temporal resolution: In Questions 4 to 6, a series of subjective experiments with a fixed spatial resolution (CIF) and variable temporal resolutions is conducted. The frame rates of the target sequence are progressively dropped from 20 fps to 1 fps as described in Section III-F. As for the first three questions (discussed in Section III-I1), the reference and target avatars were located at virtual distances of 9, 6 and 3 meters from the viewpoint, and the corresponding reference and target sequences were applied to the front surfaces of the avatars.

3) Virtual orientation vs. spatial resolution: In Questions 1 to 6, the avatars faced directly toward the viewpoint - the angle between the virtual orientation of the avatars and the camera was 180 degrees.

In Questions 7 and 8, the avatars are located at a distance of 3 meters and rotated by 30 and 60 degrees, respectively. Therefore, there is a 150 and 120 degree angular difference between the orientation of the avatar and camera.

Progressive degradation of spatial resolution was applied to the target avatar, exactly as described in Section III-I1, and the scores were recorded.

4) Virtual orientation vs. temporal resolution: A fixed virtual distance of 3 meters was chosen for Questions 9 and 10. The avatars were oriented in exactly the same way as for Questions 7 and 8. However, the temporal resolution was reduced for these questions. The frame rate of the target sequence applied to the target avatar was dropped in 10 step-sizes from 20 to 1 frame per second (Table I) and the impact of virtual orientation on the users' sensitivity to detecting the temporal resolution degradation was investigated.

5) Viewer's focal point vs. spatial resolution: In this part of the study, although two avatars were presented simultaneously to the subject, neither of them was the reference avatar. Each avatar had a red dot on the top and was labelled as a "target"avatar. One of the avatars was located closer to the view point and both had a slight rotation with respect to the avatars' directions in Questions 1 to 6.

In Question 11, while the closer avatar with constant maximum spatial resolution (CIF) was describing the instructions for the question to the subject, the spatial resolution of the more distant avatar was being degraded gradually. At the conclusion of the video, the subject was asked by the closer avatar if any change in the quality was perceived. However, the avatar to which this was referring was not explicitly indicated.

The location and orientation of the avatars were slightly modified for Question 12. In this question, the more distant avatar's video had an audio component. The avatar described the question to the subject with a different wording, and meanwhile the closer avatar's spatial resolution was reduced gradually.

The subject was allowed to complete Question 11 and 12 only once. The scores were submitted by pressing a button labelled "Yes, I have perceived a change" or "No, I have not perceived any change". The subject was also free to leave his/her comments.

Although the sequences presented in all questions were "talking head" videos, each showed a different person with a diverse variety of hand gestures.

IV. Subjective Scoring Methodology

Since the study is of the *degradation code rate* type, 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 assumed 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:

$$d_{ijk} = 5 - s_{ijk} \quad k = \{1, 2, 3, \ldots, 10\} \tag{2}$$

Then z-scores per slice are calculated:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{k=1}^{N_{ik}} d_{ijk} \tag{3}$$

$$\delta_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{k=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2} \tag{4}$$

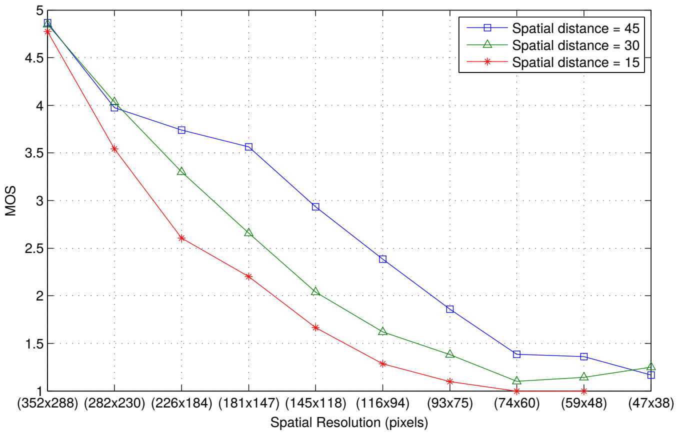


Fig. 6. Impact of virtual distance on spatial resolution

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\delta_{ik}} \tag{5}$$

where N_{ik} is the number of slices scored by subject i. The matrix z_{ij} is then 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 applied to the results to reject unreliable subjects [17]. According to this recommendation, the kurtosis of the scores is firstly calculated to determine if the scores submitted by a subject are normally distributed. The scores are considered normally distributed if the kurtosis value falls between 2 and 4. The subject is classified as unreliable if their scores are normally distributed and more than 5% of the submitted scores fall outside the range of 2 standard deviations from the mean scores. If the scores are not normally distributed, the subject is eliminated if more than 5% of their scores fall outside the range of 4.47 standard deviations from the mean scores. Since a question may be answered multiple times, first the subjects with multiple submitted scores to a given slice are identified. Then, the submitted score from that particular subject which was closest to the mean value was recognised as the answer and the rest were removed. Finally, unreliable subjects were detected and eliminated, resulting in the removal of 20 out of 233 subjects.

A. *Analysis of responses*

Fig.6 and 7 demonstrate the results of the first six questions in the subjective study. As expected, regardless of the virtual distance, recorded MOSs decline as the spatial or temporal resolutions decreases. However, the closer that the avatar is located to the camera, the greater the extent to which quality degradation is perceptible.

As shown in Fig.6 and 7, the mitigating effect of increased virtual distance on the perception of spatial resolution is greater than its effect on temporal resolution, although the effect is significant in both cases.

Fig.8 and 9 show the impact of relative avatar orientation on the viewer's perception of video quality. As described before, the avatars are rotated 30 degrees in questions 7 and 9, and 60 degrees in questions 8 and 10 with respect to the initial case (in which avatars are oriented directly toward the camera). As shown in Fig.8, the orientation has a minor impact on the viewer's perception of spatial resolution degradation and has almost no impact on degradation of temporal resolution (Fig.9). However, the skewness of scores are smaller for the higher angular states, which means more subjects have perceived the degradation in the lower temporal resolutions when the avatar is rotated farther away (Fig.10).

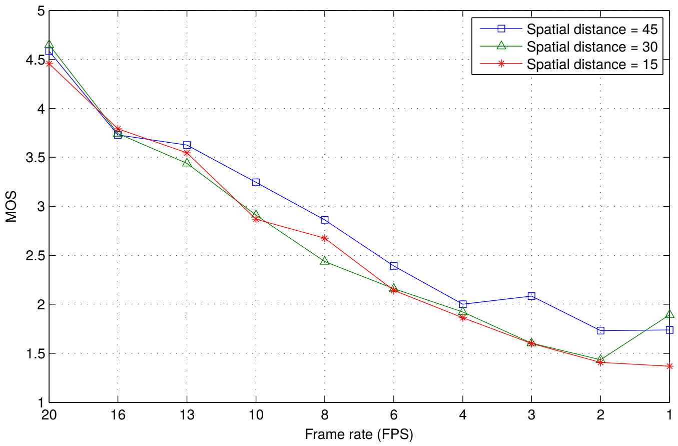


Fig. 7. Impact of virtual distance on temporal resolution

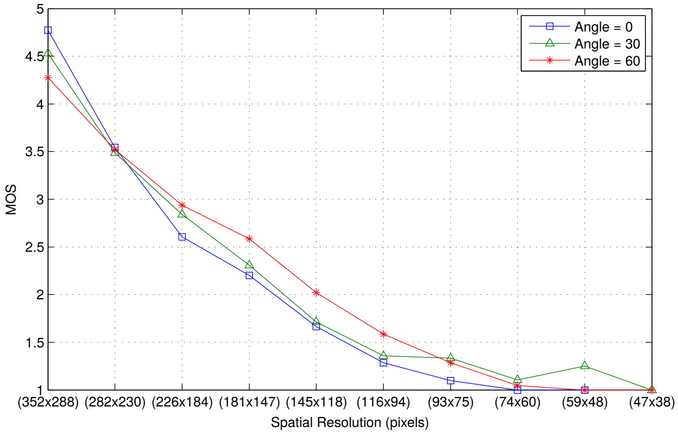


Fig. 8. Impact of orientation on spatial resolution

The reduced sensitivity of subjects to the reduction in spatial resolution when the avatar's virtual distance increased, in contrast to their relatively constant sensitivity to spatial resolution degradation as the avatar was rotated away from the viewer suggests a very close link between the projection size and shape of the video surface and the best method of deliberate video quality degradation that should be employed to render it imperceptible. The size of the video surface uniformly decreases when the avatar is moving away from the camera, while the shape of the video surface is distorted when the avatar is rotated. This observation is extensively investigated in a separate part of this research project (currently under review), in which a degradation method is proposed based on uniform size reduction of the video surface while also adapting to non-uniform situations such as rotated video surfaces.

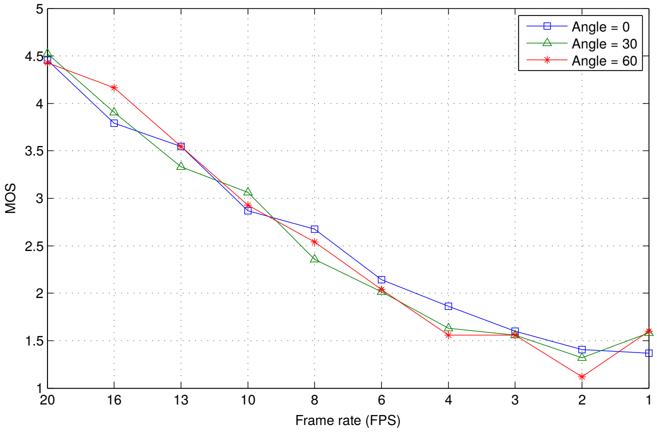


Fig. 9. Impact of orientation on temporal resolution

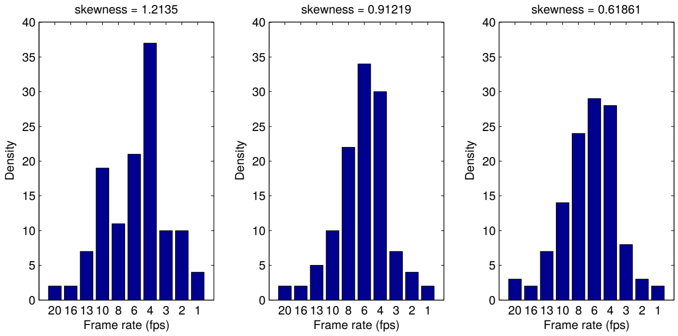


Fig. 10. Skewness of "noticeable degradation" PDF for Questions 3, 9 and 10

The influence of focal point is studied in the final two questions (11 and 12), where a video with an audio component is assigned to an avatar, and the spatial video resolution of the other avatar is degraded gradually. In the first test, the distant avatar's video is degraded, while the nearby avatar - whose video stream includes audio - focuses the attention of the subjects by describing the test to them. A similar procedure is performed in the last question, with roles of avatars reversed and the locations of the avatars slightly altered.

As demonstrated in Fig.11, 41% of the subjects participating in Question 11 did not notice any change in the quality of video, while 27% perceived some changes. However, 34% of these subject detected incorrect or unrelated quality issues, such as changes in the quality of the non-degraded video being displayed on the other avatar, lip sync and audio issues, jerkiness and frame drops. The remainder of the subjects did not submit any score for this question.

In Question 12, due to the location of avatars, 49% of the participants were able to detect the video degradation. However, 20% of the responses reported invalid issues similar to those seen amongst the responses to the previous question. 37% of the subjects did not perceive any quality change, while 14% of did not submit a response to this question. From these results, it is clear that the focus of attention plays a significant role in the perception of selective degradation of video streams.

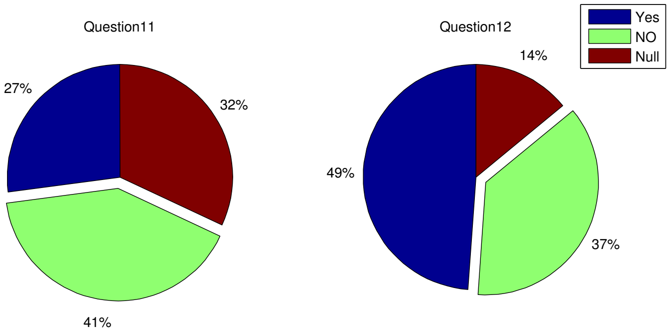


Fig. 11. Impact of focal point

B. Perceptual model of the IVC

When the 3D model of the virtual environment is mapped to a 2D display, distant avatars appear smaller than nearby avatars to the perspective projection of the scene. Hence, the dimensions (and shape) of the visible face of the avatar showing its associated video stream are dynamic, even though the size of the user's viewing window (the IVC window, i.e. the image seen by his or her virtual camera in the 3D world) 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, the avatars' projected size for different virtual distances is firstly calculated. Using nonlinear regression, a prediction model is then obtained to predict the required spatial resolution in pixels based on the virtual distance in meters. [2]

According to the submitted subjective scores, the average spatial resolution thresholds perceived as "noticeable degradation" at different distances are extracted and mapped to the modelled curve. A three-parameter exponential function is then fitted to the subjective quality scores.

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

$$s = \alpha_1 + \alpha_2 e^{(-\alpha_3 \beta)} \tag{6}$$

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 (Fig.12).

C. Bit-rate model

Many studies have addressed and modelled the impact of spatial, temporal and amplitude resolutions on perceptual quality [18], [7]. However, some of the proposed models are computationally expensive, due to the number of features and parameters used [19]. In [20], a bite rate model as a function of quantisation parameters is introduced. A mathematical perceptual model and a modelling of the bit rate in terms of the quantisation parameter and frame rate is also proposed by Wang et al. [21], [22].

[2] Note that based on the dimensions of the avatars and the virtual environment in the IVC with respect to the simulator, the values are scaled to make the simulator's data consistent with the data in the actual environment.

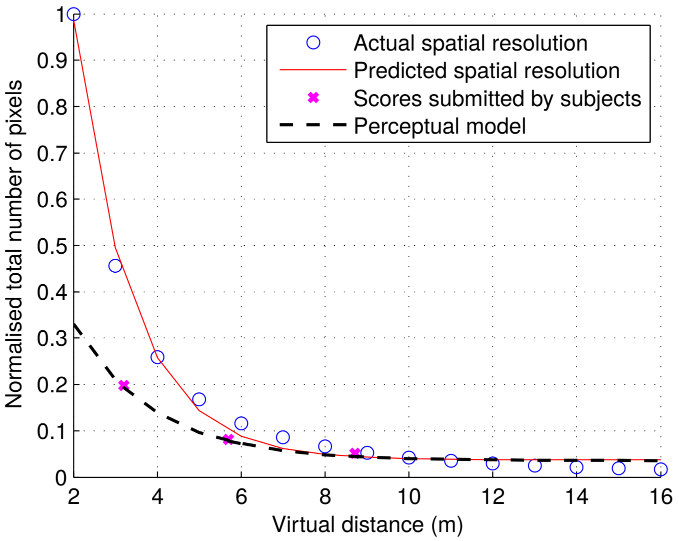


Fig. 12. Perceptual model of the spatial resolution based on virtual distance

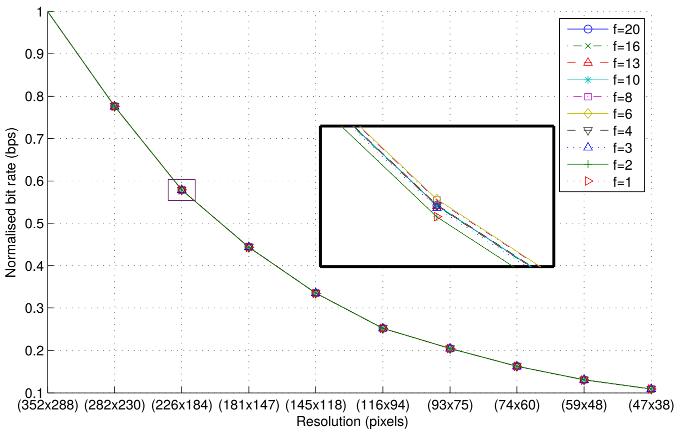


Fig. 13. Normalised rate vs. spatial resolution for different temporal resolutions

In this section, the model proposed in [21], [22] is adopted and an analytical model for video bit rate in terms of spatial and temporal resolutions is presented. In this work, the focus is on spatial and temporal resolutions, while amplitude resolution will be studied in future work. Therefore, the bit rate model is considered strictly as a function of spatial and temporal resolutions; hence the bit rate $R(s,t)$ is written as:

$$R(s,t) = R_{max} R_s(s, t_{max}) R_t(t, s_{max}) \tag{7}$$

where $R_{max} = R(s_{max}, t_{max})$ 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 the follows:

$$R_s(s, t_{max}) = \frac{R(s, t_{max})}{R(s_{max}, t_{max})} \tag{8}$$

NRS describes how the bit rate reduces 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:

$$R_t(t, s) = \frac{R(s,t)}{R(s, t_{max})} \tag{9}$$

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. The degradation mechanism described in section III-F is applied to each of the full 60 second video sequences to achieve 60 different videos with the spatial and temporal resolutions shown in Table I. 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 results achieved from all sequences are visually indistinguishable; one of the outcomes is shown in Fig.13.

The curves compiled with different frame rates overlap with each other and can be characterised by a single curve. The behaviour demonstrated in Fig.13 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.

The characteristics of the system are extensively studied in [21], where it was shown that the impacts of the functions are independent from each other, so that the normalised rate vs. quantisation parameter q and temporal resolution t can be represented by separate functions of only q and t respectively. $R_t(t)$ was shown in [21] to be a power function:

$$R_t(t) = \left(\frac{t}{t_{max}}\right)^b \quad b \leq 1 \tag{10}$$

Experimental data also confirmed the independence of s, and, as explained earlier, R_s describes the reduction of bit rate as the spatial resolution decreases. Based on the measured data, the suggested function to model the system based on spatial resolution is:

$$R_s(s) = \left(\frac{s}{s_{max}}\right)^d \quad d \leq 1 \tag{11}$$

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 processes videos frame by frame before passing the frames to the codec, the value of b was approximately 1. However, according to other studies performed with diverse video content, b has been found to vary with the intensity of motion [21]. For the videos in this study, parameter d had the value of 0.6312. After evaluating several alternative functions, we confirmed that a power function yields the minimum residual fitting error as shown in Fig.14 and 15.

Combining 10 and 11, the following overall rate model is proposed:

$$R(s,t) = R_{max} \left(\frac{s}{s_{max}}\right)^d \left(\frac{t}{t_{max}}\right)^b \tag{12}$$

where s_{max} and t_{max} are the maximum spatial and temporal resolutions respectively and should be set based on the required applications. R_{max} is also the highest bit rate achieved

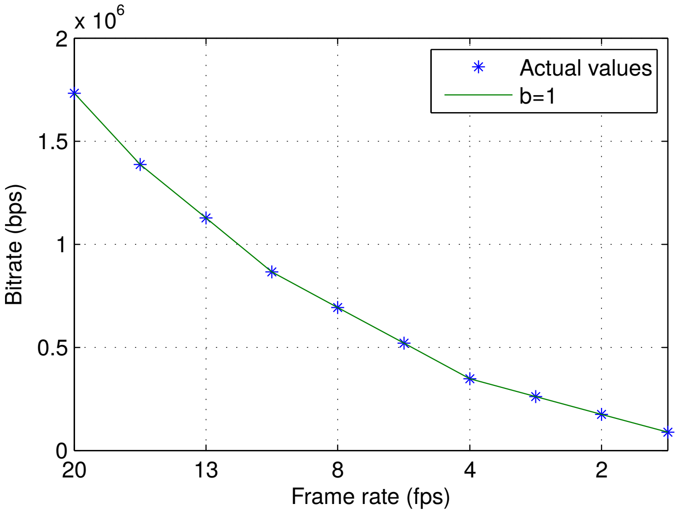


Fig. 14. Bit rate vs. temporal resolution

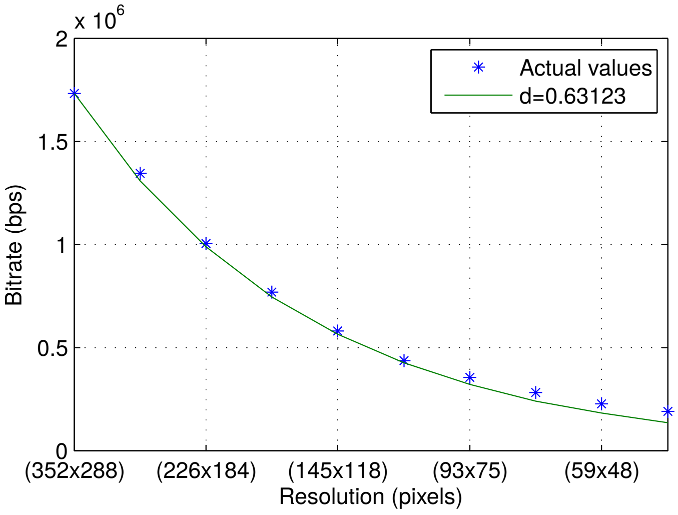


Fig. 15. Bit rate vs. spatial resolution

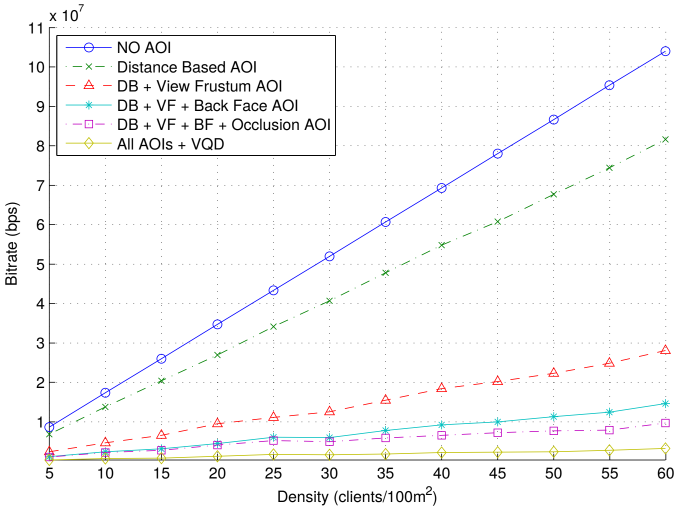


Fig. 16. Impact of density on bit rate (uniformly randomly distributed and oriented client avatars)

when spatial and temporal resolutions are set to the maximum (i.e. the resolution of the original video source) 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 is calculated. First, the bit rates for different spatial resolutions are measured with the temporal resolution set to the maximum (20 fps). In the second test, the bit rate for different temporal resolutions was calculated while the spatial resolution was fixed at CIF quality. Then, the bit rates were predicted using the proposed model; the results are shown in Fig.14. PC values of 0.9997 and 1 for predication of the bit rate based on spatial and temporal resolution was achieved respectively, showing that the model is very accurate.

V. Simulations

In previous work, several strategies for reduction of vidoe download bit rate was employed based on the *visibility* of videos from the perspective of the local client. A number of attributes was utilised to determine if the client can see a particular video. In particular if the avatar (i) is further away than a virtual distance threshold; (ii) is not within the view frustum; (iii) is facing away from the local client, with its back towards the viewer; or (iv) is occluded by another object, then its video is not needed by the local client. The combination of all of these is referred to as area of interest (AoI) management [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 12 and 6 to calculate the required spatial resolution based on virtual distance. By exploiting the model, classifying the avatars into three spatial zones based on their virtual distances to the local client and assigning different temporal resolutions to each region, the VQD mechanism was integrated into the simulator.

The IVC system in all experiments is a fixed-size (100 m×100 m) virtual room and the simulations are performed with 100 iterations in which the client avatars are placed on the floor of the environment with a uniform random spatial distribution and uniform random orientations around the vertical axis (unless otherwise specified).

A. Impact of density on bandwidth

The impact of density on bandwidth is evaluated by increasing the number of clients linearly from 5 to 60 in the IVC system. In [2], it was shown that by exploiting the AoI methods, a total bandwidth saving of 90.61% can be achieved when 60 randomly distributed clients are present in the environment. After adding the VQD mechanism, not only is any unnecessary transmission of video avoided, but also the quality of video is degraded in a way which is not perceptible to the viewer. For the simulation scenario evaluated in this section, the total reduction in bandwidth requirements is 96.85%. This is shown in Fig.16.

When the spatial distribution of avatars is changed to a two-dimensional normal distribution centred around the local

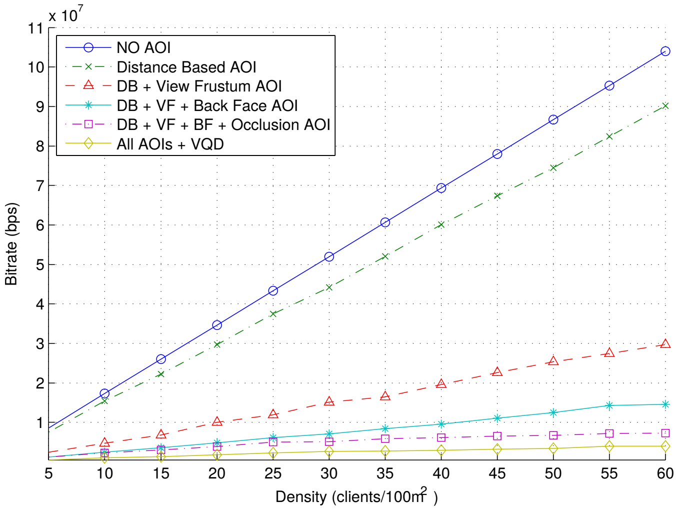


Fig. 17. Impact of density on bit rate (normal randomly distributed and uniformly randomly oriented client avatars)

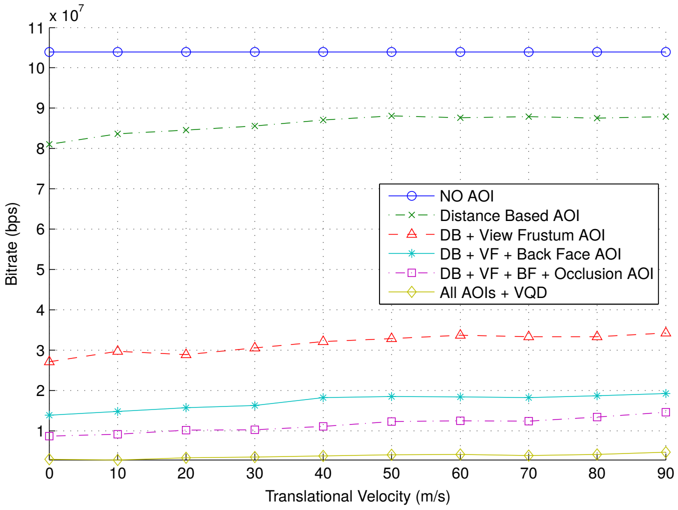


Fig. 18. Impact of translation velocity on bit rate

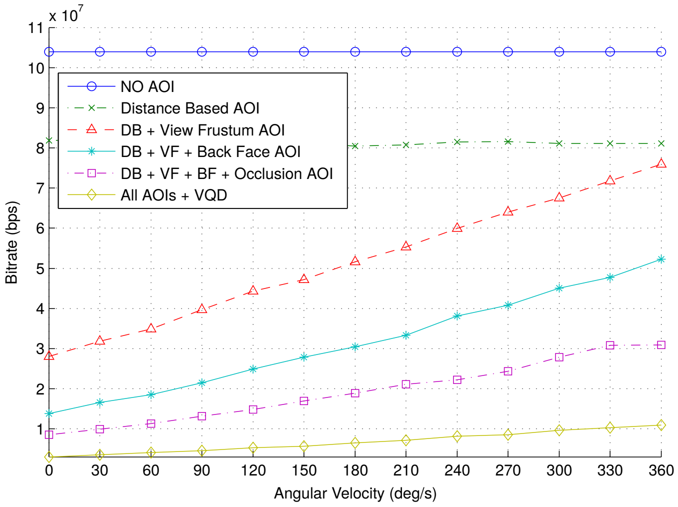


Fig. 19. Impact of angular velocity on bit rate

client with $\delta^2 = 50^2$ in both dimensions, the AoI strategies are even more effective due to occlusion culling, reducing bandwidth requirements by 92.8%. In this case, the further gains provided by VQD are smaller than for the uniform random distribution/orientation scenario, because all avatars are densely clustered around the local client and therefore require the highest video quality. However, after applying the VQD method, the bandwidth reduction increases to 96.08%. This is shown in Fig.17.

B. Impact of translation velocity

In this section, each client's translational velocity is increased from 0 to 90 m/s in steps of 10 m/s. The prediction mechanism described in [2] was utilised to predict the client's next position after the 200 ms simulated network delay.

The result shows that for this scenario, the VQD scheme reduces bandwidth requirements by an average of 67.13% compared to the use of AoI management alone when 60 clients are present in the IVC. Results are shown in Fig.18.

C. Impact of angular velocity

In this section, the angular velocities of all 60 clients in the environment vary from 0 to 360 deg/s in increments of 30 deg/s. The translational velocity is zero and the network delay is set to 200 ms.

The behaviour of the system is studied extensively in [2], where it was demonstrated that an increase of the order of 300% in required network capacity is possible when angular motion is introduced. However, after applying the VQD to the system, an average bandwidth saving of 93.54% can be achieved over the use of AoI management alone. Results are shown in Fig.19.

D. Analysing a realistic scenario

In order to simulate a more realistic scenario, both translational and angular velocities of the users are varied. 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 set to 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 ±15 degrees. The local client is placed as one of these centre points.

As demonstrated in Fig.20, occlusion culling is highly effective due to the high density of clients around the local client. Nevertheless, VQD can improve the bandwidth saving still further. For the case where the clients are clustered around 6 centre points, VQD achieves additional savings of 43.80% compared to using AoI methods alone.

E. Assessing a pathological scenario

A worst case scenario for the AoI management system would be one in which all avatars are not only in the visual range of the local client, but also in its view frustum. In

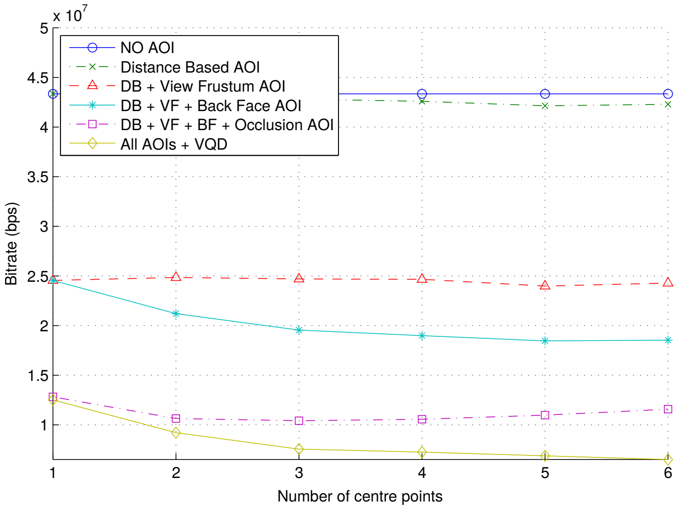


Fig. 20. Impact of clustered distribution on bit rate

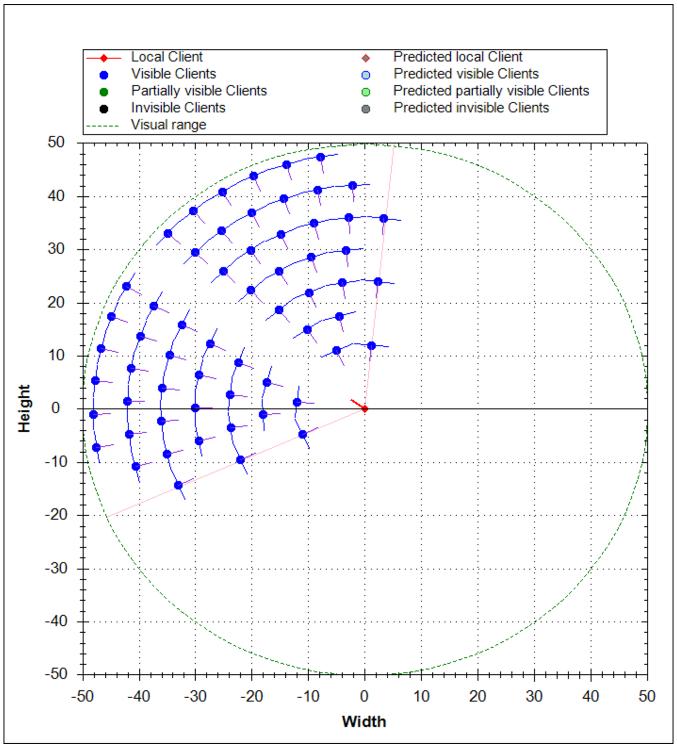


Fig. 21. Lecture theatre distribution

this scenario, all avatars face toward the local client and they are located in the 3D environment in such a way that none of them are occluded by each other or any other opaque object. Such a scenario could occur in a virtual lecture theatre environment, where all clients are arranged on a pitched floor such that those in the rear are located higher than those at the front, allowing them to see the lecturer and hence, from the lecturer's perspective, they are not occluded by each other. In this scenario, the AoI mechanism is totally ineffective and does not reduce the required network capacity requirements of the lecturer. Such a scenario is shown in Fig.21

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 Fig.22.

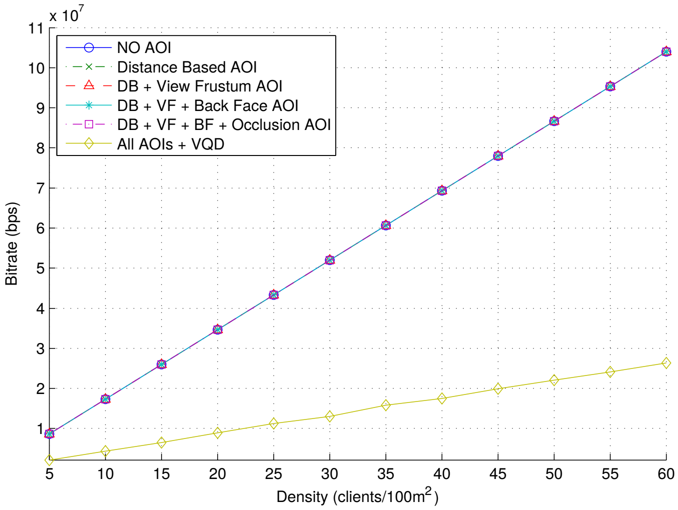


Fig. 22. Impact of density (with uniform distribution) on bit rate in a lecture theatre scenario

VI. Conclusion

In this paper, a subjective study was presented, which aimed to evaluate the impact of virtual distance and orientation on the perceptual quality of video in an immersive videoconferencing system. 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 and temporal degradation in the quality of video 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. It was shown that the bit rate of the system can be expressed as two separate functions of spatial and temporal resolutions. The rate model achieved by this key observation fits the measured rates very accurately, with an average Pearson correlation of 0.9998.

By combining the models, a complete model was developed that predicts the required spatial resolution based on the virtual distance. Using the model and categorising the avatars based on their virtual distance to three regions provides a mechanism for spatially and temporally degrading the quality of a video stream such that there is a negligible perceptual impact on the viewer. Finally, by exploiting the VQD mechanism in the simulator, many different scenarios including realistic and pathological scenarios were simulated. The results demonstrate the effectiveness of the proposed VQD strategy under a wide range of scenarios. It confirms that by using the VQD strategy, a significant bandwidth saving can be achieved in all scenarios even when the AoI mechanism is completely ineffective.

References

[1] iSeeVC Pty Ltd, “iSee,” Available URL: http://www.isee-meetings.com/ [last accessed 10 February 2016], 2014.

[2] P. Pourashraf, F. Safaei, and D. Franklin, “Distributed area of interest management for large-scale immersive video conferencing,” in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, july 2012, pp. 139 –144.

[3] ——, “Minimisation of video downstream bit rate for large scale immersive video conferencing by utilising the perceptual variations of quality,” in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, July 2014, pp. 1–6.

[4] F. Safaei, P. Pourashraf, and D. Franklin, “Large-scale immersive video conferencing by altering video quality and distribution based on the virtual context,” *Communications Magazine, IEEE*, vol. 52, no. 8, pp. 66–72, Aug 2014.

[5] G. Yadavalli, M. Masry, and S. Hemami, “Frame rate preferences in low bit rate video,” in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 1, sept. 2003, pp. I – 441–4 vol.1.

[6] J. Chen and J. Thropp, “Review of low frame rate effects on human performance,” *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 37, no. 6, pp. 1063 –1076, nov. 2007.

[7] Q. Huynh-Thu and M. Ghanbari, “Temporal aspect of perceived quality in mobile video broadcasting,” *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 641 –651, sept. 2008.

[8] A. Rossholm, M. Shahid, and B. Lovstrom, “Analysis of the impact of temporal, spatial, and quantization variations on perceptual video quality,” in *Network Operations and Management Symposium (NOMS), 2014 IEEE*, May 2014, pp. 1–5.

[9] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, “Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, T. Ebrahimi and T. Sikora, Eds., vol. 5150, Jun. 2003, pp. 198–209.

[10] N. Cranley, P. Perry, and L. Murphy, “User perception of adapting video quality,” *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 8, pp. 637–647, Aug. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.ijhcs.2005.12.002

[11] J.-S. Lee, F. De Simone, T. Ebrahimi, N. Ramzan, and E. Izquierdo, “Quality assessment of multidimensional video scalability,” *Communications Magazine, IEEE*, vol. 50, no. 4, pp. 38–46, April 2012.

[12] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, “Cross-dimensional perceptual quality assessment for low bit-rate videos,” *IEEE Transactions on Multimedia*, pp. 1316–1324, 2008.

[13] A. Eichhorn and P. Ni, “Pick your layers wisely - a quality assessment of h.264 scalable video coding for mobile devices,” in *Proceedings of the 2009 IEEE international conference on Communications*, ser. ICC’09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 5446–5451. [Online]. Available: http://dl.acm.org/citation.cfm?id=1817770.1818284

[14] W. Song, D. W. Tjondronegoro, and S. Azad, “User-centered video quality assessment for scalable video coding of h.264/avc standard,” in *16th International Multimedia Modeling Conference*, S. Boll, Q. Tian, Z. Zhang, and Y.-P. P. Chen, Eds. Chong Qing, China: Springer Netherlands, January 2010, pp. 55–65. [Online]. Available: http://eprints.qut.edu.au/30386/

[15] J.-S. Lee, F. De Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, and T. Ebrahimi, “Subjective evaluation of scalable video coding for content distribution,” in *Proceedings of the international conference on Multimedia*, ser. MM ’10. New York, NY, USA: ACM, 2010, pp. 65–72. [Online]. Available: http://doi.acm.org/10.1145/1873951.1873981

[16] I. T. S. S. S. G. 12, “Itu-t recommendation p.910, subjective video quality assessment methods for multimedia applications,” ITU, Tech. Rep., 2008.

[17] I. R. S. I.-R. S. G. . S. . B. service, “Itu-r recommendation bt.500-11, methodology for the subjective assessment of the quality of television pictures,” ITU, Tech. Rep., 2012.

[18] R. Feghali, D. Wang, F. Speranza, and A. Vincent, “Quality metric for video sequences with temporal scalability,” in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3, sept. 2005, pp. III – 137–40.

[19] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang, “Three dimensional scalable video adaptation via user-end perceptual quality assessment,” *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 719 –727, sept. 2008.

[20] W. Ding and B. Liu, “Rate control of mpeg video coding and recording by rate-quantization modeling,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 6, no. 1, pp. 12 –20, feb 1996.

[21] Z. Ma, F. Fernandes, and Y. Wang, “Analytical rate model for compressed video considering impacts of spatial, temporal and amplitude resolutions,” in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, July 2013, pp. 1–6.

[22] H. Hu, Z. Ma, and Y. Wang, “Optimization of spatial, temporal and amplitude resolution for rate-constrained video coding and scalable video adaptation,” in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, Sept 2012, pp. 717–720.