Minimisation of Video Downstream Bitrate for Large Scale Immersive Video Conferencing

Pedram Pourashraf

University of Wollongong
UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
Minimisation of Video Downstream Bitrate for Large Scale Immersive Video Conferencing

A thesis submitted in fulfilment of the requirements for the award of the degree

Doctor of Philosophy (PhD)

from

THE UNIVERSITY OF WOLLONGONG

by

Pedram Pourashraf
Telecommunication Engineering

SCHOOL OF ELECTRICAL, COMPUTER AND TELECOMMUNICATIONS ENGINEERING
2014
Abstract

Video conferencing, in particular multiparty 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.

In this thesis, an immersive video conferencing (IVC) system is introduced, which employs a 3D virtual environment to provide an intuitive space for participant to naturally interact. The real-time video streams of participants are displayed on their respective avatars in IVC. The avatars can be moved and rotated by the participants within this 3D environment, hence, the video conferencing session may include multiple simultaneous conversations (break out groups) and participants can go from one conversation to another. IVC can potentially scale to a larger number of participants provided that each participant only receives accurately adjusted videos within their field of view. This is achieved by a number of techniques developed in this research.

In the technique referred as area of interest management (AOI), transmission of unnecessary video streams to each user is avoided. The decision whether a video stream is required at a particular time would depend on the current perspective of the user. The criteria employed in this research to cull the video streams are: (i) the visual distance to the viewer, (ii) being outside of his/her view frustum, (iii) facing away, or (iv) occluded by opaque objects.

The subjective video quality assessment presented in this research show 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. Therefore, it may not be necessary for all video streams to be at the best quality or rate. Hence, another technique known as video quality differentiation (VQD) is proposed that predicts the required video quality of avatars with their respective 3D situations.

Finally, the results of VQD model are improved for any 3D situation by exploiting 3D transformation in the video quality adjustment process. The proposed perceptual pruning mechanism partitions each frame into spatial regions and calculates the required spatial resolution of each region based on their projection sizes on the screen. Consequently, a non-uniform spatially adjusted video quality can be achieved for the 3D situations that the projections of avatars are distorted.

By exploiting the proposed methods in IVC, the downstream network load for the client can be reduced up to
Statement of Originality

This is to certify that the work described in this thesis is entirely my own, except where due reference is made in the text.

No work in this thesis has been submitted for a degree to any other university or institution.

Signed

Pedram Pourashraf
22 April 2014
Acknowledgments

First I would like to thank my supervisor Professor Farzad Safaei for his helps and advice throughout my research. Simply I could not wish for a better, more supportive or more knowledgeable supervisor.

I also would like to express my appreciation to my co-supervisor Dr. Daniel R. Franklin and congratulate him for his new born daughter. Although his family needed him the most during this critical time, he helped me through my thesis to the last minute.

This work would not have been possible without the love and support of my parents, Manoochehr Pourashraf and Simin Barangizi. They made many sacrifices over the years to get me to this point. I would like to thank them here for that as well.

Finally I would like to thank my excellent friends and colleagues in iSee team and Smart Services CRC for their supports.
# Contents

List of Abbreviations ........................................... xxvi

1 Introduction ...................................................... 1
   1.1 The main research focus of this dissertation ............ 3
   1.2 Contributions of this research ............................ 3
      1.2.1 Area of Interest Management ....................... 4
      1.2.2 Video Quality Differentiation ...................... 5
      1.2.3 Perceptual Pruning .................................. 5
   1.3 Patents and publications arising from this research .... 6

2 Literature Review .............................................. 8
   2.1 Visibility determination methods ......................... 9
      2.1.1 visibility culling terminologies ................... 10
      2.1.2 Region-based visibility techniques .................. 11
      2.1.3 Point-based visibility culling ....................... 12
      2.1.4 Hierarchical Z-buffer and hierarchical occlusion map . 13
      2.1.5 Hardware accelerated occlusion culling techniques . 14
   2.2 Interest management algorithms ......................... 14
   2.3 The State-of-the-Art quality differentiation ............ 18
   2.4 Video quality assessment ................................. 19
2.4.1 Subjective video quality assessment methodologies 20
2.4.2 Visual quality metrics 22
2.4.3 Existing video quality assessments 24

3 Immersive Video Conferencing System 29
3.1 Introduction 29
3.2 Existing video conferencing technologies 29
3.3 System Overview 31
3.4 Immersive conferencing 33
  3.4.1 Uniform distribution of avatars 34
  3.4.2 Normal distribution of avatars 34
  3.4.3 Lecture theatre distribution of avatars 36
3.5 IVC Network Architecture 36
  3.5.1 Client/Server Architecture 37
  3.5.2 Peer-to-Peer Architecture 38
  3.5.3 Hybrid Architecture 39
  3.5.4 Media data streams 39
3.6 Summary 45

4 Distributed Area of Interest Management 46
4.1 Introduction 46
4.2 IVC requirements 47
4.3 AOI management algorithm 48
  4.3.1 Conservative visible set 50
  4.3.2 Conservative view field 51
  4.3.3 3D area of interest management 52
## CONTENTS

4.3.4 3D occlusion culling ........................................... 53

4.4 Simulation results ................................................ 55

4.4.1 Impact of avatar density on total downstream bitrate ...... 56

4.4.2 Impact of translational velocity on total downstream bitrate 58

4.4.3 Impact of angular velocity on total downstream bitrate .... 63

4.4.4 Analysing a realistic scenario ................................. 65

4.5 Conclusions .......................................................... 67

5 Evaluation of User Sensitivity to Video Quality Differentiation 68

5.1 Introduction ........................................................ 68

5.2 Overview of Subjective Study ................................ 69

5.2.1 Reference sequences ........................................... 71

5.2.2 Target sequences ................................................. 71

5.2.3 Subjective test design ........................................... 72

5.2.4 Presentation of test questions ................................. 73

5.2.5 Detailed description of the subjective study .................. 75

5.3 Subjective scoring methodology .................................. 77

5.3.1 Data processing ................................................... 79

5.3.2 Perceptual model of the IVC ................................. 83

5.3.3 Bit-rate model .................................................... 87

5.4 Simulations .......................................................... 91

5.4.1 Impact of density on downstream bitrate .................... 93

5.4.2 Impact of translational velocity ............................... 94

5.4.3 Impact of angular velocity ...................................... 96

5.4.4 Analysing a realistic scenario ................................. 97

5.4.5 Analysing a pathological scenario ............................ 97
6  Perceptual Pruning

6.1  Introduction ................................. 102
    6.1.1  Overview of the proposed method .............. 103

6.2  Transformation of macroblocks .......................... 104
    6.2.1  Introduction to transformations ................. 104

6.3  3D Affine Transformations ............................. 107
    6.3.1  The elementary 3D transformations ............... 107

6.4  Composition of block transformations .................... 110

6.5  Model transformation ................................ 112

6.6  View transformation ................................ 113

6.7  Perspective projection ............................... 115

6.8  Block behaviour .................................. 117
    6.8.1  Characteristics of the rendered view of the clients terminal 118
    6.8.2  Block behaviours at different virtual distances ....... 118
    6.8.3  Presentation of blocks ............................ 118
    6.8.4  Block behaviours with different virtual orientations ... 120
    6.8.5  Block behaviours with different camera orientations ... 122
    6.8.6  Characteristics of video surface ................... 126

6.9  Degradation process ................................ 131
    6.9.1  Justification of the method ....................... 132

6.10 Perceptual pruning ................................ 134
    6.10.1  The analysis of applying perceptual pruning on a static frame 134
    6.10.2  The analysis of applying perceptual pruning on a video stream 143
6.11 Simulation results ........................................... 163
  6.11.1 Impact of density .................................... 169
  6.11.2 Analysing a pathological scenario ................. 169
  6.11.3 Analysing a realistic scenario ..................... 171
6.12 Conclusion .................................................. 172

7 Conclusion ..................................................... 175
  7.1 Future works .............................................. 177

Bibliography .................................................... 179

A Experimental Settings and Subjects’ Demographics .... 190
  A.1 Introduction .............................................. 190
  A.2 Development of VQA system and experimental setup .. 191
  A.3 Subject recruitment ..................................... 191
  A.4 Human subjects training and testing ................... 192
  A.5 Flow of the study ....................................... 192
  A.6 Subjects’ demographics .................................. 193
    A.6.1 Distribution of subject by the last answered question .. 193
    A.6.2 Distribution of subject by gender .................... 194
    A.6.3 Distribution of subject by age ....................... 195
    A.6.4 Distribution of subject by eduction ................ 197

B Data Analysis of The User Study .......................... 198
  B.1 Introduction ............................................ 198
  B.2 Subjects’ contribution and votes’ distribution ....... 199
    B.2.1 Virtual distance vs. spatial resolution ............... 199
List of Figures

1.1 A screen-shot from immersive video conferencing system . . . . 2

3.1 3D representation of avatar, showing the video surface . . . . 32
3.2 The view frustum of the camera attached to the local client’s avatar 33
3.3 Avatars in an IVC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
3.4 Uniform distribution of avatars in the environment . . . . . . . 35
3.5 Clustered distribution of avatars in the environment . . . . . . . 35
3.6 Lecture theatre distribution of avatars in the environment . . . 36
3.7 Client/Server architecture . . . . . . . . . . . . . . . . . . . . . . 38
3.8 Peer-to-Peer architecture . . . . . . . . . . . . . . . . . . . . . . 40
3.9 Hybrid architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
3.10 Media streams . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
3.11 Upstream and downstream bitrate . . . . . . . . . . . . . . . . 43
3.12 Upstream and downstream bitrate (Peer to Peer) . . . . . . . 44

4.1 Visibility culling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
4.2 Conservative visible set . . . . . . . . . . . . . . . . . . . . . . . . . 51
4.3 Conservative view field . . . . . . . . . . . . . . . . . . . . . . . . . 52
4.4 Content aware ray casting model . . . . . . . . . . . . . . . . . 55
4.5 Ray casting mechanism . . . . . . . . . . . . . . . . . . . . . . . . 56
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>Density versus total bit rate normalised to the bit rate of one video stream</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>(uniform distribution)</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>Density versus total bit rate normalised to the bit rate of one video stream</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>(normal distribution)</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>Introduction of translational velocity to the system</td>
<td>60</td>
</tr>
<tr>
<td>4.9</td>
<td>Impact of translational velocity (25 avatars)</td>
<td>61</td>
</tr>
<tr>
<td>4.10</td>
<td>Impact of translational velocity (60 avatars)</td>
<td>62</td>
</tr>
<tr>
<td>4.11</td>
<td>Impact of translational velocity (60 avatars)</td>
<td>63</td>
</tr>
<tr>
<td>4.12</td>
<td>Impact of angular velocity</td>
<td>64</td>
</tr>
<tr>
<td>4.13</td>
<td>Impact of angular velocity(Percentage)</td>
<td>65</td>
</tr>
<tr>
<td>4.14</td>
<td>Clustered distribution</td>
<td>66</td>
</tr>
<tr>
<td>5.1</td>
<td>One frame from each of the reference and target avatar used in the study</td>
<td>70</td>
</tr>
<tr>
<td>5.2</td>
<td>Subjective assessment setup</td>
<td>73</td>
</tr>
<tr>
<td>5.3</td>
<td>Screen shot of the user study</td>
<td>74</td>
</tr>
<tr>
<td>5.4</td>
<td>Impact of virtual distance on spatial resolution</td>
<td>80</td>
</tr>
<tr>
<td>5.5</td>
<td>Impact of virtual distance on temporal resolution</td>
<td>80</td>
</tr>
<tr>
<td>5.6</td>
<td>Impact of orientation on spatial resolution</td>
<td>81</td>
</tr>
<tr>
<td>5.7</td>
<td>Impact of orientation on temporal resolution</td>
<td>82</td>
</tr>
<tr>
<td>5.8</td>
<td>Skewness of ”noticeable degradation” PDF for Questions 3,9 and 10</td>
<td>82</td>
</tr>
<tr>
<td>5.9</td>
<td>Impact of focal point</td>
<td>83</td>
</tr>
<tr>
<td>5.10</td>
<td>Size of video stream window (green) vs. size of the user’s viewing window</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>(blue)</td>
<td></td>
</tr>
<tr>
<td>5.11</td>
<td>Perspective projection</td>
<td>85</td>
</tr>
<tr>
<td>5.12</td>
<td>Perceptual model of the system</td>
<td>86</td>
</tr>
<tr>
<td>5.13</td>
<td>Bit rate vs. spatial resolution for different temporal resolutions</td>
<td>89</td>
</tr>
</tbody>
</table>
5.14 Normalised rate vs. spatial resolution for different temporal resolutions ........................................ 90
5.15 Bit rate vs. temporal resolution ......................................................... 91
5.16 Bit rate vs. spatial resolution .......................................................... 92
5.17 Impact of density on bit rate (Uniform Distribution) ..................... 93
5.18 Impact of density on bit rate (Normal Distribution) ................. 94
5.19 Impact of translation velocity on bit rate ................................. 95
5.20 Impact of angular velocity on bit rate ........................................ 96
5.21 Impact of clustered distribution on bit rate ............................... 98
5.22 Lecture theatre distribution ......................................................... 99
5.23 Lecture theatre distribution ....................................................... 100

6.1 Projected size of the macroblocks after 60 degree rotation about local y axis ........................................... 105
6.2 Rotation about local y axis ($\beta$) ................................................. 106
6.3 Two projected blocks, illustrating the need to consider all four corners of the projected block in order to correctly calculate the height of the projected block. ........................................ 111
6.4 Graphics pipeline ................................................................. 116
6.5 Impact of virtual distance of video surface on the blocks’ projected width ........................................... 119
6.6 Impact of virtual distance of video surface on the blocks’ projected height ........................................... 119
6.7 Presentation of blocks ................................................................. 120
6.8 Impact of virtual orientation of video surface on the blocks’ projected width at a distance of one metre ................................. 121
6.9 Impact of virtual orientation of video surface on the blocks’ projected height at a distance of one metre ................................. 122
6.10 Impact of virtual orientation and distance of video surface (distance 2 to 5 metres) ........................................... 123
6.11 Impact of virtual orientation and distance of video surface (distance 6 to 9 metres) .......... 124
6.12 Impact of virtual orientation of the camera on the blocks’ projected width at 1 metre distance .......... 125
6.13 Impact of virtual orientation of the camera on the blocks’ projected height at 1 metre distance .......... 125
6.14 Projection of video surface with camera angle zero ($\beta_0 = 0$) and 40 ($\beta_0 = 40$) ................. 126
6.15 Impact of virtual orientation of camera and distance of video surface (distance 2 to 5 metres) .............. 127
6.16 Characterisation of a video surface .................. 128
6.17 Impact of virtual orientation of camera and video surface on projection sizes of the blocks (at a distance of one metre) ............. 130
6.18 Decoder diagram .............................................. 132
6.19 DCT downsampling process .................................. 133
6.20 Perceptual pruning for various virtual orientations (2D presentation) ........................................ 139
6.21 Perceptual pruning for virtual orientations of 0 and 30 degrees (3D presentation) ......................... 140
6.22 Perceptual pruning for virtual orientations of 60 and 80 degrees (3D presentation) ......................... 141
6.23 Total bitrate of the video stream at different virtual distances ........................................... 147
6.24 Average bitrate of a video frame at different virtual distances ........................................ 147
6.25 Total bitrate of I-frames during one second at different virtual distances ..................................... 148
6.26 Total bitrate of P-frames at different virtual distances ......................................................... 148
6.27 Average bitrate of an I-frame at different virtual distances ........................................ 149
6.28 Average bitrate of a P-frame at different virtual distances ........................................ 149
6.29 Total bitrate saving achieved at different virtual distances ........................................ 150
6.30 Total bitrate saving per I-frames at different virtual distances ........................................ 151
6.31 Total bitrate saving per P-frames at different virtual distances 151
6.32 Process of calculating the perceptual quality of the video frame inside and outside the IVC system 153
6.33 Total bitrate of the video stream at different virtual orientations 156
6.34 Average bitrate of a video frame at different virtual orientations 156
6.35 Total bitrate of I-frames at different virtual orientations 157
6.36 Average bitrate of an I-frame at different virtual orientations 157
6.37 Total bitrate of P-frames at different virtual orientations 158
6.38 Average bitrate of a P-frame at different virtual orientations 158
6.39 Total bitrate saving achieved at different virtual orientations 159
6.40 Total bitrate saving per I-frames at different virtual orientations 159
6.41 Total bitrate saving per P-frames at different virtual orientations 160
6.42 Total bitrate of the video stream at different virtual orientations 162
6.43 Average bitrate of a video frame at different virtual orientations 162
6.44 Total bitrate of the video sequence at different virtual distances and orientations 163
6.45 Total and average bitrate of the sequence at different virtual orientations (virtual distance of 3 and 4 metres) 164
6.46 Total bitrate of I and P frames at different virtual orientations (virtual distance of 2 to 4 metres) 165
6.47 Average bitrate of I and P frames at different virtual orientations (virtual distance of 2 to 4 metres) 166
6.48 Total bitrate saving of the video sequence at different virtual orientations (virtual distance of 2 to 4 metres) 167
6.49 Total bitrate saving per I and P frames at different virtual orientations (virtual distance of 2 to 4 metres) 168
6.50 Impact of density on the total size of I-frames 170
6.51 Impact of lecture theatre distribution on the total size of I-frames 171
LIST OF FIGURES

6.52 Avatars with steep angular situations in IVC

6.53 Impact of lecture theatre distribution on the total size of I-frames

A.1 Distribution of subjects by the last answered question

A.2 Distribution of subjects by gender

A.3 Distribution of subjects by age

A.4 Distribution of subjects by eduction

B.1 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 1

B.2 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 1

B.3 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 1

B.4 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 1

B.5 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 1

B.6 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 1

B.7 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 1

B.8 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 1

B.9 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 1

B.10 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 1

B.11 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 2
LIST OF FIGURES

xviii

B.12 Number of subjects voted for “Not identical but hardly noticeable
degradation” at different spatial resolutions in Question 2 . . . . 206
B.13 Number of subjects voted for “Slightly noticeable degradation”
at different spatial resolutions in Question 2 . . . . . . . . . . . 206
B.14 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 2 . . . . . . . . . . . . . . . . 207
B.15 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 2 . . . . . . . . . . . . . . 207
B.16 Mean, median, mode of the submitted spatial resolutions for
“Identical quality” in Question 2 . . . . . . . . . . . . . . . . . 208
B.17 Mean, median, mode of the submitted spatial resolutions for
“Not identical but hardly noticeable degradation” in Question
2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 208
B.18 Mean, median, mode of the submitted spatial resolutions for
“Slightly noticeable degradation” in Question 2 . . . . . . . . . 209
B.19 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 2 . . . . . . . . . . . . . . . . 209
B.20 Mean, median, mode of the submitted spatial resolutions for
“Unacceptable degradation” in Question 2 . . . . . . . . . . . . 210
B.21 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 3 . . . . . . . . . . . . . . . . . . . . 211
B.22 Number of subjects voted for “Not identical but hardly noticeable
degradation” at different spatial resolutions in Question 3 . . . . 211
B.23 Number of subjects voted for “Slightly noticeable degradation”
at different spatial resolutions in Question 3 . . . . . . . . . . . 212
B.24 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 3 . . . . . . . . . . . . . . . . 212
B.25 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 3 . . . . . . . . . . . . . . 213
B.26 Mean, median, mode of the submitted spatial resolutions for
“Identical quality” in Question 3 . . . . . . . . . . . . . . . . . 213


B.27 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 3 ................................................................. 214

B.28 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 3 ................................................................. 214

B.29 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 3 ................................................................. 215

B.30 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 3 ................................................................. 215

B.31 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 4 ................................................................. 216

B.32 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 4 ................................................................. 217

B.33 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 4 ................................................................. 217

B.34 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 4 ................................................................. 218

B.35 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 4 ................................................................. 218

B.36 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 4 ................................................................. 219

B.37 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 4 ................................................................. 219

B.38 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 4 ................................................................. 220

B.39 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 4 ................................................................. 220

B.40 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 4 ................................................................. 221

B.41 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 5 ................................................................. 221

B.42 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 5 ................................................................. 222
B.43 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 5 . . . . . . . . . . 222
B.44 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 5 . . . . . . . . . . . . . . 223
B.45 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 5 . . . . . . . . . . . . 223
B.46 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 5 . . . . . . . . . . . . . . . . . . . . 224
B.47 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 5 225
B.48 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 5 . . . . . . . . . . 225
B.49 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 5 . . . . . . . . . . . . . . . . . 226
B.50 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 5 . . . . . . . . . . . . 226
B.51 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 6 . . . . . . . . . . . . . . . . . . . . 227
B.52 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 6 . . 228
B.53 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 6 . . . . . . . . . . . 228
B.54 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 6 . . . . . . . . . . . . . . . . . . . . 229
B.55 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 6 . . . . . . . . . . . . . . . . . . 229
B.56 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 6 . . . . . . . . . . . . . . . . . . . . . . 230
B.57 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 6 230
B.58 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 6 . . . . . . . . . . . . 231
B.59 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 6 ........................................... 231

B.60 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 6 ........................................... 232

B.61 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 7 ......................................................... 232

B.62 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 7 .......... 233

B.63 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 7 ................................. 233

B.64 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 7 ...................................................... 234

B.65 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 7 ...................................................... 234

B.66 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 7 ......................................................... 235

B.67 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 7 ......................................................... 236

B.68 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 7 ......................................................... 236

B.69 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 7 ......................................................... 237

B.70 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 7 ......................................................... 237

B.71 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 8 ......................................................... 238

B.72 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 8 ......................................................... 238

B.73 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 8 ......................................................... 239

B.74 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 8 ......................................................... 239
B.75 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 8 ................................. 240
B.76 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 8 ................................. 240
B.77 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 8 ........................................................................ 241
B.78 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 8 ................................. 241
B.79 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 8 ................................. 242
B.80 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 8 ................................. 242
B.81 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 9 ................................. 243
B.82 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 9 ................................. 244
B.83 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 9 ................................. 244
B.84 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 9 ................................. 245
B.85 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 9 ................................. 245
B.86 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 9 ................................. 246
B.87 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 9 ................................. 246
B.88 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 9 ................................. 247
B.89 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 9 ................................. 247
B.90 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 9 ................................. 248
B.91 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 10 248
B.92 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 10 249
B.93 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 10 249
B.94 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 10 250
B.95 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 10 250
B.96 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 10 251
B.97 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 10 252
B.98 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 10 252
B.99 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 10 253
B.100 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 10 253
List of Tables

4.1 Testing the solution convergence .......................... 58

5.1 Spatial and temporal resolutions’ step-sizes ............... 72

6.1 Objective quality assessment of differences between spatial-domain downscaling and frequency-domain image masking ............. 133

6.2 Objective assessment of reference image vs. degraded images outside of the IVC system (4 × 4 macroblocks) ..................... 135

6.3 Objective assessment of reference image vs. degraded images outside of the IVC system (8 × 8 macroblocks) ..................... 135

6.4 Objective assessment of reference image vs. degraded images inside the IVC system (4 × 4 macroblocks) ..................... 136

6.5 Objective assessment of reference image vs. degraded images inside the IVC system (8 × 8 macroblocks) ..................... 136

6.6 Zero coefficients vs. non-zero coefficients (8 × 8 macroblocks) . . 137

6.7 Zero coefficients vs. non-zero coefficients (4 × 4 macroblocks) . . 137

6.8 Objective assessment of reference image vs. degraded images outside of the IVC system (4 × 4 macroblocks) ..................... 138

6.9 Objective assessment of reference image vs. degraded images outside of the IVC system (8 × 8 macroblocks) ..................... 138

6.10 Objective assessment of reference image vs. degraded images inside of the IVC system (4 × 4 macroblocks) ..................... 139

6.11 Objective assessment of reference image vs. degraded images inside of the IVC system (8 × 8 macroblocks) ..................... 142
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.12</td>
<td>Zero coefficients vs. non-zero coefficients (8 × 8 macroblocks)</td>
<td>142</td>
</tr>
<tr>
<td>6.13</td>
<td>Zero coefficients vs. non-zero coefficients (4 × 4 macroblocks)</td>
<td>142</td>
</tr>
<tr>
<td>6.14</td>
<td>The average objective assessment results of reference frames vs. quantised frames outside of the IVC system (4 × 4 macroblocks)</td>
<td>144</td>
</tr>
<tr>
<td>6.15</td>
<td>The average objective assessment results of reference frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)</td>
<td>145</td>
</tr>
<tr>
<td>6.16</td>
<td>The average objective assessment results of quantised frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)</td>
<td>145</td>
</tr>
<tr>
<td>6.17</td>
<td>The average objective assessment results of reference frames vs. pruned frames inside of the IVC system (4 × 4 macroblocks)</td>
<td>152</td>
</tr>
<tr>
<td>6.18</td>
<td>The average objective assessment results of quantised frames vs. pruned frames inside of the IVC system (4 × 4 macroblocks)</td>
<td>152</td>
</tr>
<tr>
<td>6.19</td>
<td>The average objective assessment results of reference frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)</td>
<td>154</td>
</tr>
<tr>
<td>6.20</td>
<td>The average objective assessment results of quantised frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)</td>
<td>154</td>
</tr>
<tr>
<td>6.21</td>
<td>The average objective assessment results of reference frames vs. pruned frames inside of the IVC system (4 × 4 macroblocks)</td>
<td>154</td>
</tr>
<tr>
<td>6.22</td>
<td>The average objective assessment results of quantised frames vs. pruned frames inside of the IVC system (4 × 4 macroblocks)</td>
<td>155</td>
</tr>
</tbody>
</table>
List of Abbreviations

2D  Two Dimensional
3D  Three Dimensional
AABB  Aligned Axis Bounding Box
BFC  Back Face Culling
BSP  Binary Space Partitioning
CIF  Common Intermediate Format
CVF  Conservative View Field
CVS  Conservative Visible Set
CW  Clockwise
DBC  Distance Based Culling
DCR  Degradation Category Rating
DSCQS  Double Stimulus Continues Quality Scale
DSIS  Double Stimulus Impairment Scale
FPS  Frames Per Second
GGD  Generalised Gaussian Density
GPU  Graphics Processing Unit
HDR  High Dynamic Range
HOM  Hierarchical Occlusion Map
HSR  Hidden Surface Removal
HVS  Human Visual System
HZB  Hierarchical Z-Buffer
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IVC</td>
<td>Immersive Video Conference</td>
</tr>
<tr>
<td>JND</td>
<td>Just Noticeable Distortion</td>
</tr>
<tr>
<td>MMOG</td>
<td>Massively Multiplayer Online Games</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MVC</td>
<td>Model View Control</td>
</tr>
<tr>
<td>NAT</td>
<td>Network Address Translator</td>
</tr>
<tr>
<td>NRS</td>
<td>Normalised Rate vs. Spatial Resolution</td>
</tr>
<tr>
<td>NRT</td>
<td>Normalised Rate vs. Temporal Resolution</td>
</tr>
<tr>
<td>OC</td>
<td>Occlusion Culling</td>
</tr>
<tr>
<td>PC</td>
<td>Pearson Correlation</td>
</tr>
<tr>
<td>PSNR</td>
<td>Peak Signal to Noise Ratio</td>
</tr>
<tr>
<td>PVQMs</td>
<td>Perceptual Visual Quality Metrics</td>
</tr>
<tr>
<td>PVS</td>
<td>Potentially Visible Set</td>
</tr>
<tr>
<td>QP</td>
<td>Quantisation Parameter</td>
</tr>
<tr>
<td>QS</td>
<td>Quantisation Step-size</td>
</tr>
<tr>
<td>SDSCE</td>
<td>Simultaneous Double Stimulus for Continuous Evaluation</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>SSCQE</td>
<td>Single Stimulus Continues Quality Evaluation</td>
</tr>
<tr>
<td>SSL</td>
<td>Secure Sockets Layer</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>VF</td>
<td>View Frustum</td>
</tr>
<tr>
<td>VFC</td>
<td>View Frustum Culling</td>
</tr>
<tr>
<td>VIF</td>
<td>Visual Information Fidelity</td>
</tr>
<tr>
<td>VQD</td>
<td>Video Quality Differentiation</td>
</tr>
<tr>
<td>VQM</td>
<td>Video Quality Model</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Due to the growth of the Internet and availability of more bandwidth over lower-latency connections for the end users, multimedia applications are becoming increasingly popular. Multimedia conferencing over the Internet is one the applications in demand. Audio conferencing was introduced as the first generation of multimedia conferencing, where the system provides a shared aural space between the participants. The visually augmented audio conferencing system was introduced later, in which a shared visual space is provided in addition to the aural space. However, the visual space in this system is usually limited to texts, images, graphics and in some cases some forms of animation. Finally video conferencing systems were introduced, which enables the participants to communicate not only via audio, but also by using live video. Hence, the visual space is extended and supports live video streams of the participants. Although, the quality of video streaming has improved in recent years, video conferencing systems have not changed fundamentally since their introduction.

Existing video conferencing applications are limited in conference size. This lack of scalability is partly technical, due to limitations of resources such as bandwidth and processing power, and partly due to their presentation and the limitations of social protocols with large groups.

A novel video conferencing system is considered in this research that not only overcomes the scalability issues, but also revolutionises video conferencing. Im-
**Introduction**

**Figure 1.1** A screen-shot from immersive video conferencing system

**Immersive video conference** (IVC) is a 3D virtual environment designed for video conferencing. All participants exist in the environment in the form of a simplified visual representation of the user - known as an **avatar** - with the user’s webcam video stream displayed on the front surface of the avatars. IVC allows a large number of simultaneous users to interact and communicate effortlessly in a shared virtual environment using spatially accurate audio and video, where, in essence, the real life characteristics of a human gathering are being emulated (Figure 1.1). IVC represents a combination of positive attributes of video conferencing systems and distributed virtual environments. The participants have freedom to move within the environment, join other discussions or form a separate conversation group. The combination of rich visual and aural scenes creates a sense of immersion and provides a comfortable space for users to naturally interact. Moreover, the natural human behavior of ‘mingling’ in a crowd becomes possible, where participants have peripheral awareness of surrounding conversations and dynamically move from one to another.
1.1 The main research focus of this dissertation

The key challenge for the development of IVC is scalability, that is, the ability for a large number of geographically dispersed users to meet and interact efficiently. For a video conferencing application that supports multi-person-to-multi-person interactions, the number of required audio and video streams transmitted over the network grows as the square of number of participants. This means that the application’s requirement with respect to network capacity cannot scale economically. Consequently, most current solutions impose a rather modest upper limit on the maximum number of participants. This scalability barrier applies both to audio and video streams, but given that video is more bandwidth hungry, this research focuses on video scalability.

By fusing together the concepts of distributed virtual environments and video conferencing, not only is the sense of immersion and usability enhanced, but also scalability is improved. The addition of the third dimension means that not all avatars would necessarily fit within the visual range of a given user. This mimics the character of crowd interactions in the real world and is an acceptable constraint for the users. Even within the visual range of the user, not all avatars will have the same size and shape, and hence, may not require the same video quality and resolution. Again, as our user study confirms, this variation of quality based on perspective is in accordance with users’ expectations. These characteristics are exploited in this dissertation to improve the scalability of IVC.

1.2 Contributions of this research

The main contributions of this thesis are summarised below.
1.2.1 Area of Interest Management

The first technique to manage scalability is discussed in Chapter 4 and is referred to as area of interest (AOI) management, in which the distribution of participants’ videos is dynamically altered in response to changes in the AOI of users. In other words, every user will only receive those videos that are required for his/her current view in the 3D environment. Area of interest management is achieved by exploiting four different techniques; In the first step, the video streams of the avatars that are located beyond a given visual distance are discarded; this technique is known as distance-based culling. Next, view frustum culling is applied to the surviving avatars. In this process, the avatars outside of the local avatar’s view frustum are omitted from the list of required video streams. Back face culling additionally prevents transmission of videos of avatars inside the view frustum but facing away from the local client. Finally, occlusion culling removes video streams of avatars occluded by static and dynamic opaque objects from the required streams' list.

Although the AOI technique effectively reduces the required downstream bitrate of the system, it greatly increases sensitivity to movements of avatars. Both translational and rotational movements of participants will change the visual scene and consequently the distribution pattern of video streams. Without any mechanism to compensate for this effect, users will see ‘blank’ avatars when changing perspective (e.g., turning around) or when a new person enters their view frustum. Hence, a prediction mechanism is introduced to prevent this effect. In this approach, the composition of the future view frustum is constantly predicted based on current movement patterns, and using a control feedback loop, the video stream distribution pattern is adjusted to prepare the clients for the change of view. In an actual interactive scenario, the additional bandwidth usage due to this prediction is quite modest. With judicious application of the AOI method with the prediction mechanism in a crowded virtual space, the total downstream bitrate of the system are typically reduced by 90%.
1.2.2 Video Quality Differentiation

The 3D nature of IVC also helps scalability in another way. Avatars within one’s AOI typically have different virtual distances and orientations. Since perspective projection is employed in IVC, avatars with different virtual situations have different projection sizes on the screen. In this thesis, it is shown that by differentiating the video quality of avatars based on their 3D situations, not only the required network capacity can be decreased, but also the quality reduction would not be noticeable. A user study was conducted to evaluate the users’ sensitively to the change of avatars’ video quality at different 3D situations. A web-based mean opinion score (MOS) video quality assessment was designed to study the impact of virtual distance and orientation on the perceived spatial and temporal video quality degradation. The study contained 12 questions, in which the first six questions analysed the impact of virtual distance on the perceptibility of spatial and temporal video quality reduction, while questions six to ten studied the influence of virtual orientation on the perceived video quality. The last two questions investigated the effect of focal point on the subjects. A total of 233 individuals participated in the study and the results show that subjects are less sensitive to video quality degradation when the virtual distance and the skewness of avatar orientation increase. However, the impact of virtual distance found to be more pronounced than virtual orientation. This behaviour is the result of non-uniform change of avatar’s projection size in the case of rotation. Using the survey results, a video quality differentiation model was proposed as two separate functions of spatial and temporal resolutions. The outcomes confirm that the model predicts the required resolutions based on the virtual situation of avatars accurately.

1.2.3 Perceptual Pruning

The survey revealed that uniform degradation of video quality as exists in the current video quality differentiation models (e.g. hierarchical video coding or multiple description coding) is inefficient in some 3D situations (e.g. orientation change). To address this inefficiency, a novel model is proposed, which incor-
porates 3D transformations in the degradation process. The aim of *perceptual pruning* is to develop a suitable technique for IVC, so that individual video streams can be pruned before transmission to the client based on their exact 3D situations. In this model, the spatial regions of a video frame are not necessarily degraded uniformly and they may have different levels of degradation based on their projection sizes. In the process of video quality adjustment, the surface to which the avatar’s video is assigned, is partitioned based on the video frame’s macroblocks’ partitioning, and the perspective projection of each block is calculated. Next, the spatial resolution of the video frame is adjusted based on the projected blocks’ sizes on the screen. Since this process is applied on the server side (or peers) before sending the streams to the clients, it should be computationally feasible. Hence, a DCT downsampling method is proposed that only requires entropy decoding of the frames. Using this method, a frequency mask is applied to the video data in frequency domain to reduce the spatial resolution of the frame accordingly. By exploiting perceptual pruning in addition to AOI mechanism, total downstream bitrate can be further improved by 56.58% with a negligible perceptual impact.

### 1.3 Patents and publications arising from this research

The IVC system used in this research has been trialled successfully in Australia and it is now in the process of commercialisation. Hence, the intellectual properties of this research’s contributions are protected by our industrial partner. *Area of interest management* technique developed in this research is one of the contributions patented and publish as follows:


- 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. [2]

The video quality differentiation model is also secured under the same patent. The proposed perceptual pruning mechanism is the subject of another patent, which has been filed recently. Due to the legal issues regarding the protection of the intellectual properties, submission of the papers related to these contributions are delayed. However, two journal papers are under the process of submission.


Chapter 2

Literature Review

Increasingly, users regard the Internet as a meeting place, where they can form communities and interact with groups of people as part of their work, play, education, or social interaction with family and friends. This phenomenon is likely to create a significant demand for multiperson-to-multiperson communications. Virtual worlds such as Second Life [6,7] support avatar-based rich communications, a wide range of online activities, 3D content creation, and development of various in-world teaching and learning tools [8].

On the other hand, video conferencing applications facilitate interactive video and audio communication between individuals. However, the conventional video conferencing systems cannot scale to support a large number of participants. The common practice of displaying videos of participants as rectangular tiles on the screen, the so-called Brady-Bunch model, is very restrictive and cannot scale. By showing the ‘relevant’ participant only, such as the one with audio activity, the scalability is improved but often resulting in strained social protocols and cognitive fatigue for the users.

In this research, the cognitive problem is addressed by combining the concepts of video conferencing and the positive attributes of distributed virtual environments, where the video of participants are shown on the front surface of their respective ‘avatars’. Similarly to other virtual environments, each participant is represented by an avatar and can roam freely in a 3D space. However, unlike
graphical avatars, the participants’ avatars in an IVC display their real-time video and their voice travels in the virtual environment in accordance to the expected properties of the propagation of sound.

The concept of IVC is novel and no prior work was found in this area. Furthermore, to achieve such a system, the video of each participant is broken down into frames and each frame is applied as a texture on the front surface of the avatar in the 3D environment (refer to Section 6). Hence, the related works in the area of 3D projection and visibility determination, existing interest management techniques and video coding and assessment of video quality are presented in this chapter.

This chapter is structured as follows: first, the 3D characteristics of the system is analysed in Section 2.1 and the existing visibility determination mechanisms in computer graphics are investigated; then, the interest management methods used in other contexts like gaming is studied in Section 2.2 and the reasons that they are not applicable in the context of IVC are expressed; and finally, the video quality differentiation and assessment techniques are presented in Section 2.3 and 2.4.

2.1 Visibility determination methods

In computer graphics, visibility determination is a challenging problem. Due to the huge size of the data set in a typical 3D environment, real-time approaches to minimise the load of graphics pipeline are essential.

Visibility culling techniques usually exploit two basic methods: Back face and view frustum culling [9]. Back face culling methods avoid rendering of geometries that are facing away from the view point, while view frustum culling techniques avoid rendering of geometries out of the viewing frustum.

Low-level back face culling is adopted in graphics processing units (GPUs) nowadays and this feature is accessible in the majority of graphics programming APIs such as OpenGL and Direct3D. The higher-level methods such as hierarchical
back-face culling [10] can also be exploited alternatively.

The view frustum culling methods, which is usually implemented at object level can be enhanced by the hierarchical data structure. Although the view frustum culling methods can efficiently reduce the number of rendered objects, the reduction is highly dependent on the viewer’s position and scene’s structure. In addition, the more complex occlusion culling technique prevents rendering of geometries that are occluded by other opaque objects.

Current techniques for visibility determination can be classified as follows: Point-based algorithms such as ray casting [11], which perform computation with respect to the current position of the view point; and region-based algorithms that perform bulk computation in a specific region, which is valid anywhere in that region [12].

In another classification, Object precision methods perform visibility computation on raw objects [13], while Image-precision algorithms like Hierarchical Z-Buffer [14] utilise the discrete representation of objects during the rasterisation process. In object-precision methods, computation can be performed by CPU and there is no need for transmission of data between CPU and GPU and, hence no dependency on graphics cards, while Image-precision methods may benefit from graphics card’s capability as hierarchical occlusion map (HOM) [15] does. Finally, some algorithms take advantage of the characteristics of architectural interiors in which scenes are constructed from cells (rooms) and cells are connected to each other using portals (doors and windows), known as cell and portal methods [16]. These techniques are mostly used in region visibility culling, while other methods are more generic and can be applied to all scenes.

2.1.1 visibility culling terminologies

Regardless of the type of the visibility culling methods, the objects that affect the techniques can be classified as occluders and occludees; the objects that hide other objects or the ones that are hidden. Since, the ultimate goal of occlusion culling is improving the rendering frame rate, the required computation time
is attempted to be minimised. Hence, most algorithms construct a *potentially visible set* (PVS), which is a superset of the exact set of visible objects and may be a small number of invisible ones [17], [18]. In this process, some occluded objects may be included in the set. However, no visible object is missed.

### 2.1.2 Region-based visibility techniques

Most of the works in this category take advantage of the characteristics of architectural scenes. For instance, scenes can be divided into cells that link through doors and windows called portals. The fundamental concepts of PVS and *densely occluded environments* are implemented and introduced by [17–19]. In all these works, PVS is usually preprocessed for each cell and used during run-time. To cache the PVS information and reduce computation, Teller [20, 21] and teller and Séquin [18] proposed both 2D and 3D approaches. In 2D method, during the processing stage, scene is divided into cells by detecting major opaque objects such as walls using a binary space partitioning (BSP) tree. Other opaque objects are ignored in this step. Then non opaque objects on the cell boundaries are determined and tagged as portals. Accordingly, an *adjacency graph* is constructed by the information. This graph identifies which cells are linked directly through the portals. It is obvious that an object in a cell can be visible from another cell, if there is at least one portal in between the cells. In order to determine more accurate cell-to-cell visibility, existence of sightlines that can connect a point from one cell to another is tested. Moreover, the PVS can be further culled by cell-to-cell visibility culling based on view volume of the observer. In this test, a cell is determined visible, not only if the cell and at least one portal linking cells are in the view volume, but also a sight line exists between cells [22]. Although Teller extended his technique [21] for 3D space by utilising a parameterisation of line space, performance of the method is not acceptable and provides a weak visibility in 3D space [20].
2.1.3 Point-based visibility culling

The Point-based objects precision methods usually rely on big occluders or cells- and-portals concept. A point-based cell-and-portal method is proposed in [23], which performs a recursive depth traversal of the cells based on the projections of the portals on the screen instead of preprocessing the PVS for each cell. The proposed algorithm provides a real-time conservative culling method. In the culling process, the cell which contains the view point is rendered and the portals inside the view frustum of the viewer are detected. Next, the portals are overestimated using the axial 2D bounding box of the projected vertices of the portals. Then, the cells linked through the portals are determined. Although the same procedure is applied to the adjacent cells, the new portals are clipped against the old ones to construct a smaller visible “window”, until no visible portal remains.

The method proposed in [24], benefits from visual events, which limits the visibility tests of occludees to only when their visibility statuses are changed. In this technique, a small set of visual events are constructed and the events are tracked, as the user moves and the visual relationship among objects changes. In order to minimise the number of visual events that needs to be tracked, the silhouette edges and vertices of the relevant primitives are constantly checked.

As discussed before, image-precision methods operate at the image level. They work with discrete representation of image. Hence Point-based image-precision algorithms usually exploit the fact that the image can gets filled up fast during the rendering. Hence, the subsequent object are culled with no extra computation. Generally, image-precision techniques are easier to implement, due to the discrete array of finite pixels in the image. Furthermore, in scenarios that the scene is constructed by many small object with no large occluders, these algorithms have a big advantage over object precision techniques. Since, the projection of small primitive can be accumulated on the image, the culling process using graphic rasterising hardware is performed more efficiently.

Ray casting is an example of point-based image-precision methods. In order to
construct the rendered image, a ray is emitted from the eye through the pixels toward the scene. The first object, which is intersected by the ray defines the content of the pixel. The main advantage of this method is that the occluded objects are never rendered. However, ray casting is usually very expensive in terms of computation. For instance, in a naive implementation of the method, the intersection of each ray with all the geometries in the scene should be calculated. However, in more advanced methods, the objects are sorted back to front and then intersection calculation is performed. Hence, the calculation can be terminated very fast [11].

2.1.4 Hierarchical Z-buffer and hierarchical occlusion map

Hierarchical Z-Buffer (HZB) is developed based on (HSR) algorithms, in which the visible portion of the primitive in a scene is determined at image-level. However, the hierarchical Z-Buffer uses two hierarchies: an octree in object-precision and a Z-pyramid in image-precision level. Z-pyramid is a layered buffer which contains the content of the Z-buffer at the best resolution layer. Subsequent levels are constructed by halving the resolution of the previous level in all dimensions.

The constructed octree is traversed top-down front-to-back and each node is checked for occlusion, if any node is detected as occluded, the test is terminated, otherwise, its children are tested continuously. Finally all primitives related to a non-occluded leaf are rendered. To decide if a node is visible, the nearest Z-value of the face is compared to the coarsest level of the pyramid and the comparison repeated hierarchically to the finest level until the object’s visibility can be determined.

The hierarchical occlusion map (HOM) is very similar to the HZB but optimised for computation with graphics hardware. The main enhancement compared to HZB is decoupling the visibility test into an overlap test and depth test [15]. HOM stores the opacity information of the scene as a black and white map, in which the white area represents the projection of the occluders on a black background. The Z-values are collected separately. The coarser levels of this
map are created by averaging squares of 2×2 pixels of the finer level. Then overlap test starts by projecting the bounding box on the scene and finding a level in the hierarchy, where pixels have almost the same size with the projected box. If the box overlaps just black pixels, it means that the box cannot be occluded, otherwise it needs to be further analysed via a depth test.

2.1.5 Hardware accelerated occlusion culling techniques

Based on the feedback loop of graphics cards, which makes it possible to detect any change in the z-buffer, hardware vendors made occlusion culling a built-in feature. By exploiting this feature, rendering of a complex model can be prevented by first checking if the model is potentially visible. The rendering performance can significantly be improved by only checking the bounding boxes of complex object for visibility before rendering the actual objects [25].

Although exploiting this capability can enhance the culling performance, a naive approach that checks every single object for occlusion slows down the graphics card and hence the rendering frame rate. For instance, applying such an implementation to the scene, in which all the polygons are projected in a back to front order, reduces the hardware performance significantly, as none of the objects would be occluded and all objects are checked at each frame.

Studies in [26] overcomes this problem by ordering the primitives front to back before the occlusion culling, which can potentially minimises the number of required hardware queries.

Parallel occlusion culling, minimising the required memory for updating the z-values [27] and hierarchical z-buffer [14] are other techniques to improve hardware accelerated occlusion culling.

2.2 Interest management algorithms

One of the application of interest management is in the context of gaming. Unlike traditional games, which are typically limited to around 16 players,
massively multiplayer online games (MMOG) allow hundreds of players to play simultaneously together in a persistent world [28]. The game actions (e.g. movements and picking up objects) are known as game state information. Scaling for MMOG is a serious issue, since broadcasting all the state information to all players in the game is not feasible.

On the other hand, the view of the world should be consistent in order to provide a shared sense of space among players. As a solution, players can maintain a copy of game states on their local machines. However, when each player affects the world, the game states of all other players should be updated. Consequently, the number of messages sent over network to keep all the states consistent grows exponentially as the number of players increases. In order to overcome this challenge, massively multiplayer online games exploit interest management techniques to restrict their communications to only relevant state information. The relevant states are the ones which affect the player (e.g. the objects that can be seen or are near). In other words, interest management is the process in which the relevance of each piece of state information to each player is determined [29]. The state information relevant to a player, usually corresponds to the perception of the player’s avatar in the world. some techniques simply consider the radius around the player as the region of interest, while more advanced ones consider obstacles in the world as well.

Interest management (IM) methods can be classified into space-based (extrinsic) and class-based (intrinsic) categories [29]. The relative position of the player’s avatar in the world determines the relevant states in space-based interest management techniques, while class-based methods are determined by attributes of an object (e.g. type of the object). The aura-nimbus information model [30] is the cornerstone of the space-based interest management. Aura-nimbus is based on proximity; the aura can be defined as the area of space which bounds the object, while the nimbus or area of interest is the space in which an object can perceive other objects. In a simple model, the aura and nimbus can be represented by fixed sized circles around the object. Then Object $a$ is interested in Object $b$, when the circle around Object $a$ (nimbus) intersects the circle around
Object \textit{b} (aura). MASSIVE-1 is a purely aura-nimbus based model introduced in [31]. The aura-nimbus based model is enhanced in [32] by using standard message oriented middleware (MOM) technologies to scale the message dissemination of MMOGs.

Although pure aura-nimbus based models can be granular sufficiently to prevent any irrelevant state information from being sent to players, the high computational cost of calculating the intersection between the nimbuses and aurás of objects can become a bottleneck for the scalability of the system [33]. In order to reduce the computation, region-based methods are used in many studies as an approximation [34, 35]. In the region-based models, the world is initially divided into static regions. The interest management technique then obtains the regions that intersect with the area of interest of the player (subscriber) and constructs the area of subscription from the union of the intersected regions. The area of subscription is then used as the new area of interest of the player. Although this approximation is cheaper in terms of computational complexity, the performance of the model is highly dependent on the size and shape of regions. The square tiling mechanism is popular and straightforward, however, hexagons are more efficient in approximation of the area of interests [35]. On the other hand, designing regions with arbitrarily shape or size is possible with systems like Spline [36]. Discussed models are not environment/terrain aware and do not take into account obstacles in the world. However, visibility based interest management techniques consider the vision of players instead of a circle around them. The visibility-based model proposed in [37] is called RING. In this client-server model, the environment is partitioned into rectangular regions and real-time visual interaction between regions is supported by precomputed visibility between regions. Hence, when a state information is changed, only the clients that can potentially perceive the change receive the corresponding message.

On the other hand, the visibility-based interest management presented in [38] is a distributed model, which retrieves the visibility state of objects from each client’s visibility culling process performed in the course of graphic rendering.
Although in this techniques, the determined visibility is precise at no cost, the assumptions of the model is; First the position data are provided and; Second the result of visibility culling performed by the renderer is accessible. Thus, in the systems that the position information is the main source of messages or the visibility culling information is not available, this model is inapplicable. Furthermore, in this work, the position data is distinguished form the visual data and it is claimed that with no need of extra calculations, the interest management of visual data can be extracted from the visibility culling results. Since the visibility determination is already performed by the renderer. However, it is not considered that majority of visibility culling techniques ignore dynamic occluders (see Section 2.1) due to the cost of computation. Hence, the proposed model is inefficient, if the major cause of occlusion is dynamic objects. Moreover, the network delay involved in sending the retrieved data to other clients in the distributed model is neglected, which may cause discrepancy between the clients’ view of the world.

Recent works are basically enhancement of the mentioned methods, by introducing features like dynamic cell-based [39] or adaptive [40] interest management techniques.

The Performance of different interest management algorithms incorporating various level of visibility and map conformance is studied in [41]. Moreover, a real massively multiplayer game is utilised in this study and the measurements are achieved by employing both human and computer-generated game actions. Finally, the efficiency of the interest management techniques are compared in terms of number of state information messages sent between players.

Unlike, the presented studies in this section, which aim at filtering out the irrelevant game state information, our approach in Chapter 4 is not concerned with state information but strives to minimise the total video downstream bitrate of immersive video conference(IVC) (definition of IVC is available in Chapter 3). Due to the fundamental difference of game state information and video streams, totally different mechanisms are required for area of interest management. For instance, the state information of an object (e.g. remote player) located vir-
tually close to the player but outside of his/her field of view is considered as relevant game state. Since, the object may affect the player’s state information (i.e the local player can be shot by the remote player from behind). In another case, even if area of interest of the player is constructed based on his/her visibility, the object, which is in the view field but facing away should be recognised as relevant. Since, state information of the object can be changed by interaction (i.e. local player can shoot the remote player from behind). However, none of these cases are valid for area of interest management of IVC. In fact, in the presented simulations in Chapter 4, the performance of region-based mechanism is compared against our approach. The results show that the downstream bitrate saving of distance based method is insignificant (refer to Section 4.4).

2.3 The State-of-the-Art quality differentiation

Current models for video quality differentiation either use hierarchical video coding (HVC) [42] or multiple description coding (MDC) [43]. In both models, the video stream is split into a number of sub-streams, called layers or descriptions in HVC and MDC respectively. The user who receives all the sub-streams will be able to decode the video at the maximum possible quality. If some sub-streams are not sent by the source or dropped by the server/network, the receiver will decode a lower quality video. The primary difference between HVC and MDC is that in HVC the layers are dependent on each other. The user must receive a base layer to be able to decode the video at all. The other layers are enhancement layers that would improve the decoded video but without the base layer cannot be decoded individually. In contrast, the multiple descriptions of MDC are independent of each other and can be decoded in isolation. This flexibility, however, comes at the cost of higher bit rate for the sub-streams. For our purposes, this distinction is not relevant and the following discussions will apply to both.

HVC and MDC were both designed for a multicast video distribution scenario where there is a heterogeneous population of recipients with different video
quality requirements (e.g. a mobile phone versus a TV screen) or bandwidth capacity (wireless versus high-speed landline). Unfortunately, these techniques do not address the needs of an IVC video quality differentiation system for the following reasons:

- Each enhancement layer or description sub-streams is concerned with improving the overall video quality equally in every part of the video. In IVC systems, the quality differentiation may not be uniformly distributed over the spatial extent of a video. For example, the video may be partially occluded or the avatar may be rotated such that some parts of the video require lower quality due to perspective variations.

- The number of layers or descriptions is usually quite small as it is computationally expensive for the source to produce and manage many different layers. This means that the resulting quality differentiation is rather coarse. However, in IVC we require finer granularity for quality adjustment based on numerous factors, such as virtual distance, angular orientation, etc.

### 2.4 Video quality assessment

Due to the significant growth of communications infrastructure and network technologies, more number of computers and digital systems are connected by networks such as the internet. This connectivity causes an enormous flow of data into homes and businesses. Majority of this data is in the form of digital visual signals such as images and videos. Consequently many products and applications are developed upon visual signals (e.g. IPTV, video conferencing application, YouTube). These services have gained a huge attention and now evaluation of their quality of service (QoS) is a challenge.

As human eyes are the final receiver of these signals, assessment of the signals should be preferably performed by humans. Undoubtedly human subjects are the best judge for the perceptual quality of the visual signals. However, subjective assessment of the signals have some shortcomings as well. Subjective
assessment is expensive and time consuming, due to the requirement of laboratory equipments and repetitive viewing sessions. Furthermore, it may not be possible to conduct the study in some situations such as real-time signal manipulations. Finally, the study can be affected by physical conditions, emotional states and personal experience. Hence, objective assessments are also developed that aims at constructing computational models to predict perceptual quality of the signal. In this section, assessment methodologies, visual quality metrics and existing subjective and objective video quality assessments are presented.

2.4.1 Subjective video quality assessment methodologies

Video technologies aims at reproducing the world as seen by human eyes. Thus, human visual system (HVS) and the way that it represents the world should be understood first. This process involves with understanding the characteristics of the visual system, limitations of visual perception, sensitivity of the visual system to temporal and spatial changes and many more. In [44], not only the fundamentals of the visual system and the HVS-based video coding approaches are introduced, but also the subjective and objective assessment methods, quality metrics and test procedures are discussed. In addition, international recommendations for subjective video quality assessment has provided detailed guidelines on how to organise and conduct a video quality experiments in ITU-T Recommendations P.910 and P.911 and ITU-R Recommendation BT.500-11. Double stimulus and single stimulus test methods are explained in the recommendations.

In the case of a single stimulus test, sequences are displayed without an explicit reference and subject are required to rate the quality of each individual sequence. On the other hand, in a double stimulus test the sequences are displayed pairwise. The reference sequence is first shown to the subject and then an impaired or degraded version of the same sequence is displayed. After each pair of sequences, the subjects are asked to provide their ratings for the quality or change in quality between the impaired and the reference sequence. There are variety of single and double stimulus tests and each of them has its own ad-
vantages. For instance, The double stimulus continues quality scale (DSCQS) is claimed to be less influence by the context of the video, while single stimulus continuous quality evaluation (SSCQE) is considered to achieve more representative quality estimates for quality monitoring applications. In DSCQS, pairs of video sequences are displayed in a randomised order to the subjects. Furthermore, each pair is shown twice and the subjects are requested to rate the quality of each sequence in the pair. Hence, context effects such as severity and ordering of the impairment are more tolerated by the subjects and the memory-based biases from previously viewed sequences are minimised. In the standard double stimulus methods, the video sequences are typically 10 seconds long and the test only provides a single quality score for each video sequence.

On the other hand, by exploiting the SSCQE method, the quality of an arbitrarily long video sequence can be rated by a slider mechanism with quality scales. Moreover, this method provides subjective score at a higher sampling rate. However, it might decreases the accuracy of the scores comparing the double stimulus methods. Subject may also be influenced by contextual factors. In addition, the subject might put too much attention to the movement of the slider in the right direction and loose the track of the absolute slider position in the rating system. In such a case, the viewer’s scores might drift over the course of the test. More detailed information regarding the subjective video quality assessment methodologies and their advantages and disadvantages is provided in [45].

These methodologies have been adopted and utilised in different studies. However, new variations of them are also proposed. For instance, another form of double stimulus testing called the simultaneous double stimulus for continuous evaluation (SDSCE) is introduced in [46], where the reference and impaired sequences are presented side by side. In some works, the recommendations are totally modified to better suit their requirements. Not only the methodologies specify the order and the maximum duration of the sequences (typically between 10 and 15 seconds), but also viewing conditions (e.g. viewing distance, monitor resolution, display brightness and contrast) are defined in these
methodologies, which can contradict the study’s purpose. As an example, a work is presented in [47], which intents to assess the quality of experience of IPTV and video on demand services in real life environments. Staelens et al. believe that the controlled environments as explained in the recommendations do not reflect a consumer’s natural environment of watching television. Unlike a user study’s condition that people primarily concentrate on evaluating the audiovisual quality in a limited time, in a typical scenario people usually lean-backward and watch television with friends or family for a longer period. Hence, a new methodology known as full-length movie quality assessment is proposed that enables assessment of videos under the same condition that users typically watch television.

In this research, video quality assessment in the context of IVC is presented in Chapter 5. Due to the novelty of IVC and presentation of it for the first time by us, the existing methodologies can not satisfy our requirements. Hence, the double stimulus methodology is adopted and modified to better suit IVC’s requirements. In the utilised method, not only the reference and degraded videos are paired, but also they are presented in a 3D environment side by side. In order to increase the sample rate of the study, the degradation level of the impaired sequences is changed over time and subjects are asked to rate the change of quality by pressing the corresponding buttons. Moreover, a web-based mean opinion score (MOS) video quality assessment is designed that not only mimics the IVC’s output, but also enables more number participation in the study. The recommended wordings are also changed to better express the situation. Details of the conducted user study is explained in Section 5.2.

2.4.2 Visual quality metrics

How to evaluate visual signals plays a big role in developing dependent technologies such as acquisition, watermarking, compression, transmission and presentation of visual signals. Hence, different visual quality metrics are proposed. However, availability of the reference signal differentiates the metrics. If the evaluation requires the complete reference signal then it is categorised as a full
reference (FR) metric. If partial information of the reference signal is sufficient, the metric is one of the reduced reference (RR) ones and finally if the signal is evaluated without any prior knowledge of the reference signal, the metric is classified as a non-reference (NR) metric.

Mean square error (MSE), signal to noise ratio (SNR) and peak signal to noise ratio (PSNR) are few examples of the full reference metrics. These simple metric measure the signal fidelity [48] and they are widely accepted and used. However, these metrics cannot accurately represent the perceptual quality of signals, especially when the noise is not additive [49]. Hence, the perceptual visual quality metrics (PVQMs) are introduced. PVQMs are the objective computational models that are developed based on the properties of human visual system to predict human perceived visual quality [50]. The structural sensitivity of human eyes is considered in SSIM [50] and its relatives [51,52] to predict perceived quality of image.

Furthermore, the contrast sensitivity function and the contrast masking effect, which are two HVS characteristics are utilised in a wavelet domain decoupling algorithm in [53] to improve MSE and PSNR. The outcome separates detail losses and additive impairment for image quality assessment and effectively works for majority of distortion types. The visual information fidelity (VIF) measure in [54] is achieved by quantifying two mutual information quantities. In this procedure, the mutual information of the reference image between the input and output of the HVS channel is quantified versus the mutual information of the test image between the input of the distortion channel and the output of the HVS channel. VDP [55], PDM [56] and PEVQ [57] are examples of HVS-based metrics for video coding. Maximum distortion that human eyes are incapable of perceiving it, is called just-noticeable distortion (JND). JND has been utilised in image assessment. In order to extend the estimation of JND for video, temporal and spatial properties of HVS should be taken into account. A JND estimator for video in DCT domain is proposed in [58], which incorporates spatio-temporal contrast sensitivity function. MOtion-based Video Integrity Evaluation index, or MOVIE index [59] not only integrates both spatial and temporal aspects of distortion, but also exploits 3D Gabor filter banks to
adaptively guide spatio-temporal filtering by using optical flow estimation. The scores achieved by MOVIE index have a close correlation with human perceived quality and the algorithm outperforms the algorithm developed and submitted to the video quality expert group FRTV Phase 1 study [60].

Due to the partial existence of the reference signal, the reduced reference (RR) quality metrics work either based on *modeling image distortion, modeling HVS* or *modeling natural image/video statistics*. In the first approach, video degradations caused by specific distortions such as MPEG-2 compression can be gauged [61]. While, HVS properties are considered in the second approach and several HVS related features are extracted to detect edge information changes, spatial information losses, contrast information and color impairment. By using these factors an efficient metric known as video quality model (VQM) was developed that performed statistically better than other models submitted to the VQEG FR-TV Phase 2. Hence, VQM was standardised by ANSI and included in Draft Recommendations from ITU-T Study Group 9 [62]. In the third approach, the metrics rely on the fact that image statistics can represent real-world distortions. For instance, generalised Gaussian density (GGD) is utilised in [63] to depict wavelet coefficient distribution. Moreover, better performance is achieved in [64] by employing GGD to depict the coefficient distribution in recognised DCT domain.

Developing non-reference (NR) quality metrics is extremely complicated. Thus, evaluation of image/video with specific distortions is studied in majority of available works. The image/video statistics for implementing NR metrics are discussed in [65–67]. An extensive study of visual metrics is provided in [68]. Due to the excellent correlation of SSIM with human perceived visual quality and acceptance of MSE and PSNR in existing works, these metrics are used in our study to evaluate performance of perceptual pruning method in Chapter 6.

### 2.4.3 Existing video quality assessments

The impact of video distortions caused by compression and transmission technologies on the perceptual quality of video is presented in many works. How-
ever, the study in [69] not only led to a publicly available video quality database known as LIVE, but also presented a large-scale subjective video quality assessment conducted on the database, which contains 150 distorted video sequences. Degraded videos are achieved by applying four different distortion processes including MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks, and through error-prone wireless networks on ten high quality videos of natural scenes. Each sequence is assessed by 38 human subjects and the recorded difference mean opinion scores were demonstrated. Finally the performance of several full-reference (FR) video quality assessment algorithms on the new database was analysed and presented.

New formats of visual signals such as 3D image/video, high dynamic range (HDR) images, tablet and mobile videos are recently delivered to more number of users. Hence, evaluation of these contents has become a hot research topic. For instance, quality of experience of video streaming on tablet devices is studied in [70]. In this work, 216 video sequences encoded with H.264/AVC are corrupted by typical wireless channel transmission errors and displayed to 40 subjects. The study is conducted on Apple iPad 2 and Samsung Galaxy Tab GT-P1000. In the collected MOS scores, degradations caused by packet loss and high playout delay were perceived as the worst and best qualities respectively. Among the metrics used to evaluate the study objectively, SSIM had the highest correlation with the subjective scores.

Due to the recent popularity of stereoscopic 3D contents and availability of required hardware for home users, in [71] two subjective assessment datasets are presented to evaluate efficiency of several different 3D video coding techniques in terms of transmission and perceived quality of stereoscopic 3D videos. In the first study, the performance in a lossless transmission channel is studied, while the second study investigates the error concealment. Wang et al. demonstrated that temporal degradation of stereoscopic 3D videos has a big impact on the perceived quality and the result is not acceptable, while spatial downsampling of the content may lead to a better bitrate efficiency. Furthermore, multiview
video coding (MVC) performed better than H.264 simulcast coding in this work.

Moreover, a new database of stereoscopic 3D videos in Full-HD captured by a semi-professional 3D camera is presented in [72]. The shooting conditions of stereoscopic 3D sequences vary from close and medium to long range distances. Furthermore, the usual television contents and 3D effects are included to be utilised in wide variety of applications such as subjective assessments. Finally several common types of spatial and coding degradations as well as enhancements are applied to the sequences and a subjective assessment is conducted. The outcome shows the uniform representation of video quality spectrum and confirms that the proposed evaluation is well balanced.

The objective assessment of other newly emerged visual signals is studied in [73]. However, Ma et al. believe that there are still many challenges involved in designing accurate objective quality metrics and developing corresponding applications. Due to the complexity of the metrics and the number of components such as HVS modelling and signal statistical prosperity involved in the metrics, they cannot simply be utilised in multimedia processing applications for the newly emerged visual signals.

Although quality assessment of videos in 3D immersive environment is presented in this research, the concept of 3D is totally different from stereoscopic 3D videos and they should not be mistaken. The objective of presented work in Chapter 5 and 6 is to find the minimum required video downstream bitrate by differentiating quality of different spatial regions of videos to enhance the scalability of the system.

In most existing studies on subjective quality assessment of scalable video coding, temporal, spatial and amplitude (i.e., signal-to-noise ratio (SNR), usually controlled by the quantisation step-size (QS) or quantisation parameter (QP)) resolution are studied separately or jointly. Temporal resolution 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 aforementioned factors [74].

Temporal resolution in combination with SNR is studied in [75] and [76]. It is proven that a high frame rate is perceptually preferred to a high frame quality for content with fast motion. It is also shown that for slow motion content, the frame rate reduction has a minor perceptual impact on subjects.

The impact of spatial and amplitude resolution was also studied in [76]. The reference image has a spatial resolution of 320×192 pixels, and the frame rate was fixed and equal to 30 Hz. Three further 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 the paired comparison methodology was conducted in [77] to investigate the trade-off between the spatial resolution and temporal resolution. This study utilised MPEG-4 encoding and the spatial and temporal resolution of video stimuli varying from 40-100% of QCIF resolution at between 5 and 25 frames per second. 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 [78–81], 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 5.3.2), rather its size changes based on the virtual distance and orientation of the viewer relative to the visible avatar. Moreover, we focus on the joint impact of spatial or temporal resolution with respect to
virtual distance and orientation as described in Section 5.3.3. 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 of the prior works have utilised bit rate and perceptual quality models to find the best spatial and temporal resolution to achieve minimum bit-rate, which is the key objective of Chapter 5. A comprehensive overview of subjective quality assessment of scalable video coding is available in [82].
Chapter 3

Immersive Video Conferencing System

3.1 Introduction

In recent years, with progressive increases in Internet bandwidth, high-quality multiparty video conferencing systems have become widespread, and are now considered a valuable collaborative service. However, the majority of existing tools are limited in the maximum conference size that they can support. This research presents a novel approach to video conferencing, which not only can overcome the major technical limitations of scalability, but also represents a significant new type of immersive experience for the participants.

In this Chapter, Section 3.2 first outlines existing video conferencing technologies and their shortcomings, Section 3.3 gives an overview of the IVC system, Section 3.5 then defines the IVC’s network architecture, and finally Section 3.6 summarises and concludes the Chapter.

3.2 Existing video conferencing technologies

The immersive video conference (IVC) in this research is a multiparty video conferencing system which employs a common 3D virtual environment to create a sense of immersion and provide a comfortable space for users to naturally
interact.

The video conferencing tools and virtual environments have been used widely for different applications. However, they have never been combined as presented in this research to overcome the issues as outlined below:

- Existing video conferencing systems\(^1\) do not scale to support a large number of simultaneous users. This lack of scalability is partly technical, due to limitations of resources such as bandwidth and processing power, and partly due to strained social protocols and increased cognitive load associated with dealing with many simultaneous audio and video streams.

- The network and computing resources required to support high quality video conferencing is significant. Most solutions require designated rooms that are fitted with specialised hardware. Apart from the cost of furnishing such rooms, which are often underutilised, there is the additional inconvenience of booking and travelling to these venues for the users.

- In a typical video conferencing session, some participants may be physically colocated, while others are remote. This dichotomy of situational context can create barriers in establishing an inclusive communication atmosphere and cause some of the participants to feel excluded.

- In a practical session, the size of conference is small and usually a strict protocol of *one person talking at a time* is observed. Multiple simultaneous conversations, such as breakout groups or workshop-style discussions, are not directly supported by the technology. This is mainly due to the 2D presentation of the conference, in which the screen is either divided equally amongst the videos of participants or they are represented as small thumbnails and the speaker’s video is maximised in the middle. In the first approach, the size of each video diminishes as more users join the session, while in the second approach, dynamic changes to the central video, which would occur as people enter and leave the discussion, are

\(^1\)To the best of our knowledge, the video conferencing systems like Skype, GoToMeeting and WebEx have a restriction of maximum 4 to 15 concurrent participants in a conference.
confusing and distracting. Moreover, the conference is limited to a single conversation as the same audio stream is received by all the participants.

In contrast, IVC systems utilise a novel three dimensional (3D) design which can scale to a large number of participants and create a unified context for communication. The technology is easy to use and only requires access to a computer with a webcam and microphone. Although the users may be in different locations, the presence of the participants in a common 3D environment provides a comfortable space for social interaction. Furthermore, the methods proposed in this research address some of the technical problems outlined above and potentially greatly increases the maximum the conference size. This will be extensively discussed in Chapter 4, 5 and 6.

3.3 System Overview

The IVC is a 3D virtual environment designed for video conferencing. All participants exist in the environment in the form of a simplified visual representation of the user - known as an avatar - with the user’s webcam video stream displayed on the front surface of the avatar. This surface is known as the video surface, as shown in Figure 3.1.

The avatars can rotate about their vertical axis and move around freely in the 3D environment. This environment is arbitrary; it could be a photo-realistic replica of an existing site or a a simple visual cue to provide a common spatial perception among the users and facilitate the natural “mingling” of participants. The voice of each participant emanates from the avatar and propagates in accordance with the expected properties of sound travelling within the virtual environment. As such, the voices arrive at the listener from the right direction and distance, and the aural scene experienced by each user is an accurate representation of the soundscape that would have been heard while in the presence of a real crowd. The combination of rich visual and aural scenes provides an engaging experience for users to interact and communicate with each other.
The system can scale to support a very large number of participants, both technically and socially, and can also facilitate sharing of multimedia content to enrich the experience. A short video of the *iSee* IVC system designed by our research team, which is used as the basis for the investigations in this thesis can be found in [83].

As a result of the distributed characteristic of the IVC system, each participant has a field of view within which the avatars of other users may be visible. This is achieved by attaching a virtual camera to the local avatar’s object in the world. The camera has a *view frustum*, which is the region of space in the world that appears on the screen. The exact shape of this region varies depending on what kind of camera lens is being simulated, but typically it is a frustum of a rectangular pyramid (Figure 3.2). The number and identity of avatars within the view frustum will dynamically change as a result of movement of avatars within the environment.

In Figure 3.3, the avatars are shaped like a floating display board that can be moved and rotated in response to their owners’ commands (i.e. key presses and
Additional information may also be shown on the front surface of the avatars. In the example of Figure 3.1 and 3.3, the participants’ name and a voice activity indicator are visible. The back surface of the avatar, in the current implementation, can be used to display any information desired, such as the name of the participants or an image of their business card. Of course, this particular shape is just one possible form to represent real-time video of participants in the environment and other forms can easily be implemented to suit the requirement of a particular application scenario.

### 3.4 Immersive conferencing

Since, participants in an immersive video conference are free to roam (rotate and move) in the 3D virtual environment, any type of distributions of avatars might be formed in the environment. However, some type of avatars’ distributions are observed more frequently and hence analysed and exploited in this research to minimise the required network capacity to support the service. These formations are usually based on the social protocols exist in real-life scenarios, which are adopted in virtual worlds.
If the number of avatars per unit area (one square metre) of the immersive environment is called *virtual density*. The downstream/upstream bitrate of the system at high virtual densities with different type of distributions is studied in this research.

### 3.4.1 Uniform distribution of avatars

When participants in the IVC system are not instructed explicitly or implicitly to form groups for a specific purpose (e.g. study groups) and if the virtual density of the immersive conference is increased, the participants keep a comfortable space between their avatars and others. Consequently, the avatars are not positioned too close or too far from each other and hence a uniform distribution of the avatars is achieved (Figure 3.4).

### 3.4.2 Normal distribution of avatars

In more realistic scenarios, avatars usually form conversations and hence clusters in the environments. Based on the nature of the conversation and the role of each participant in the conversation, a normally distributed formation of avatars about one or more centre of attentions (the centre of attention can be an avatar (e.g. the speaker)) is constructed (Figure 3.5).
Figure 3.4 Uniform distribution of avatars in the environment

Figure 3.5 Clustered distribution of avatars in the environment
3.4.3 Lecture theatre distribution of avatars

As explained in Section 3.3, due to the 3D characteristics of the system each avatar has a field of view and not necessarily all other avatars are within the visual range of the viewer. However, there might be situations that not only the visibility of all avatars in an avatar’s visual range (Avatar A) is required, but also all other avatars need to see the avatar (Avatar A) or a certain part of the environment. Such a scenario could occur in a virtual lecture theatre environment, where all avatars 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 (Avatar A) and hence, from the lecturer’s perspective, they are all visible and not occluded by each other. Such a scenario is shown in Figure 3.6.

The impact of the proposed mechanisms at different virtual densities with mentioned distributions are analysed in Chapter 4, 5 and 6.

3.5 IVC Network Architecture

In this section, the network architecture of IVC systems, including hardware, software and communication protocols are discussed. The network architecture of IVC systems can be classified into two main categories: Client/Server archi-
tecture and Peer-to-Peer. Each design is used for specific purposes as described in the following sections.

3.5.1 Client/Server Architecture

Client/Server system are constructed so that the requested services (e.g. sharing of media data or distribution of video streams) from the IVC clients can be fulfilled by powerful dedicated servers.

Although a desirable quality of service can be provided using this architecture, the cost of service and maintenance of the servers are high. Thus this architecture is more suitable for large organisations and corporate environments.

In the current implementation of the iSee IVC, the server architecture consists of three main servers as shown in Figure 3.7. The authentication server is responsible for validating the clients’ identities. The clients' information is encrypted and saved in a database on the authentication server. Clients connect to the server using secured connection over HTTPS. Secure sockets layer (SSL) creates a secure connection between the client and authentication server, over which the confidential user information can be sent securely. When a user is authenticated, a list of sessions is shown to the user. Then based on the selected session, encrypted internet protocol (IP) addresses of control and media servers are sent to the client. Then a user datagram protocol (UDP) flow is initialised between the corresponding servers and the client to exchange the state and medidata.

During a session, clients upload their state information (e.g. position and orientation) to the control server and it broadcasts the state data to all clients in the session. At the same time, clients upload their video and audio data to the media server, which multicasts the packets among the other clients. However, audio and video packets are treated differently. In the next three chapters a set of realtime strategies are proposed to minimise the amount of transferred video data between media server and clients. Hereafter the word “server” refers to the media server, except where explicitly mentioned otherwise.
Scalability of the system is enhanced by the control server, which has a load balancing functionality. It monitors the network traffic and adds new media servers if required to shift the traffic to the new servers. If there are multiple servers available for a session, the clients are connected to a server which is geographically closer to them by measuring the average round trip time. These strategies are out of the scope of this research.

### 3.5.2 Peer-to-Peer Architecture

A Peer-to-Peer (P-to-P or P2P) network is a distributed architecture can be adapted for IVC purposes. In this model, the IVC participants share a fraction of their own hardware resources (such as processing power, network capacity) and the function of the media and control servers are transformed to local processes on the clients’ machines (Figure 3.8). However, the authentication server still acts centrally and verifies clients’ identities. It is also possible for the more powerful clients to support weaker clients by performing some computationally
intensive tasks for them. This capability is provided through the scalability enhancement of the distributed control process.

Although the Peer-to-Peer structure scales well, the most significant drawback of this model is that the shared resources are necessary to provide the service offered by the IVC. Hence they should be accessible by other peers directly, without passing intermediary entities. For example, peers that are behind a proxy or a network address translator (NAT) might not be accessible by other peers, with the result that the isolated peer fails to connect to the conference.

### 3.5.3 Hybrid Architecture

In order to overcome the restrictions of the Peer-to-Peer model, a hybrid network architecture is proposed, where not only clients share their resources to enhance the service, but also central servers are provided to avoid any loss of the service to isolated clients.

In this model, the control and media servers maintain their roles, as in the Client/Server architecture. However, powerful clients are also identified and tagged as *power peers*.

The power peers can act as the servers and multicast the state, video and audio data of weaker peers amongst other peers. In such a scheme, the bandwidth and processing capacity of peers are determined through a test phase and according to the results, a power peer can be given a certain fan-out to serve the weak peers. In addition, the scheme needs a routing technique that forwards the data from the weak peers to the power peers and from there to the destination peers.

Based on this mode, the system can scale with lower cost, while the isolated peers can still be served.

### 3.5.4 Media data streams

Video conferencing tools are amongst the most bandwidth-hungry applications, and management of the required network capacity to support these services is
Figure 3.8 Peer-to-Peer architecture
Figure 3.9 Hybrid architecture
a challenging problem. The majority of traffic is usually created by the media streams, of which the video components are the largest portion. Transmission of media data consists of *upstreaming* and *downstreaming*. When clients/peers send their media data including video and audio to the server or other peers, it is called upstreaming and when they receive other clients media data from the server/peer, it is known as downstreaming (Figure 3.10).

The amount of network upstream capacity that a client requires to send his/her video stream to the server/peer is called the *upstream bitrate per client* and the amount of network downstream capacity which needs to be consumed by a client to receive only one stream of video from the server/peer is called the *downstream bitrate per client*. The total amount of network downstream capacity that a client needs to receive all required video streams is called *total downstream bitrate*. In this research, minimisation of total downstream bitrate is studied (Figure 3.11).

In the Client/Server model, clients only upstream their own media data to the server, while they are required to downstream all other clients’ media data from
the server. In the traditional multiparty video conferencing systems in which this is the case, the total downstream bitrate of each client grows linearly in proportion to the number of participants in the conference, while the required network capacity of the whole system increases in proportion to the square of the number of participants. This expansion is the main reason for the limitation of existing systems in handling a large number of concurrent participants.

On the other hand, clients in a peer-to-peer model need to act like a server and usually send their media data directly to all other peers. This causes the upstream traffic load on the clients to be affected by the number of participants as well (Figure 3.12). Regardless of the network architecture, the proposed mechanisms in this research are aimed at further reducing the required network capacity of IVC systems.
Figure 3.12 Upstream and downstream bitrate (Peer to Peer)
3.6 Summary

This chapter presents the IVC system, a novel video conferencing technology that delivers video and audio streams to 3D environments. IVC systems enable more natural social interaction between the participants than is possible with a conventional 2D conferencing system. Furthermore, the immersive environment can scale to a larger number of participants. This is primarily because it offers the ability to choose the video streams that the participant is interested in via an intuitive social protocol, and secondly because bandwidth and processing resources can be allocated according to the viewer’s field of view in the 3D environment. The poor scalability of existing multiparty video/audio conferencing systems, which is largely due to the quadratically increasing demand for network capacity on the server side and with increasing numbers of clients can be solved through the use of an IVC system. In Chapter 4, a range of techniques referred to as area of interest (AOI) management are proposed to decrease the required bandwidth and hence increase the number of potential concurrent participants in the IVC environment.
Chapter 4

Distributed Area of Interest Management

4.1 Introduction

The major bottleneck for multiparty video conferencing systems is the required network and server transmission capacity to support the service, which in the case of a traditional video conferencing system grows in proportion to the square of the number of participants.

_Distributed area of interest management_ is one of the techniques, which is proposed in this thesis to reduce total downstream bitrate and improve scalability by avoiding the transmission of videos of invisible avatars. Restricting data transmission to the video stream of the visible avatars may be achieved by the combination of a number of techniques. These include view frustum culling (VFC), back-face culling (BFC) and occlusion culling (OC) algorithms. We have intentionally used the same terminologies as common in computer graphics. However, it must be noted that the nature of algorithm and its application will be different in the context of IVC. This chapter aims to investigate the application of the _visibility culling_ techniques in IVC and demonstrate the amount of capacity saving achieved by the proposed methods. The contributions of this chapter are as follows:

- Adapting the existing culling techniques to the requirements of IVC and
developing a novel method to avoid transmission of unnecessary video streams.

- Designing a prediction mechanism based on velocity, orientation of clients and network delay.

- Introducing a conservative visible set and view field to handle the feedback delay.

This chapter is structured as follows: in Section 4.2 and 4.3 related works and IVC requirements are discussed and the proposed method is introduced; Section 4.4 presents simulation results and analysis; and finally Section 4.5 summarises the conclusions of the research.

### 4.2 IVC requirements

Although the techniques discussed in Section 2.1 use similar terminology to what we will introduce for IVC, there are fundamental differences, which make them inapplicable to an IVC. First, the visibility culling techniques in computer graphics focus on reducing the data rate for the graphics pipeline and improving rendering efficiency. However in this study the aim is to reduce network transmission capacity. Second, most studies on visibility determination need to deal with millions of primitives, hence dynamic occluders are mostly ignored. By contrast in this system, AOI needs to be applied on avatars that are dynamic objects and one of the main causes of occlusion in the system. Finally, in the case of existing culling methods, data is transmitted between CPU and graphics card, which has negligible latency, but in the proposed IVC model, data should be sent to server or other peers over the network. Thus network delay must be considered.
4.3 AOI management algorithm

Since IVC is a distributed system and each avatar is allowed to move and rotate independently, the shared visual space is perceived differently from each avatar’s perspective. The avatar, which is under the control of the local client and its view is projected on the screen of the participant is called local avatar. On the other hand, the avatars, which exist in the environment, but they belong to the remote participants with their video streams on the front surface of the avatars, are called remote avatars or simply avatars. For simplicity of understanding and presentation, the proposed algorithm is initially introduced in a two dimensional (2D) situation, and then extended to a three dimensional (3D) environment. However, in a 3D environment with a flat floor, the 2D algorithm may also be applicable. Note that, the procedure, in which the outcome of a method or algorithm is determined is referred as a test. In Figure 4.1, the local avatar’s view point is represented by a red dot and the orientation of the avatar is indicated by a red arrow. The visible remote avatars are represented by a solid blue line and dashed black lines indicate the invisible remote avatars. First, based on the maximum visual range of the local avatar, a cull pattern based on distance is applied and avatars out of a specific distance from local avatar are eliminated from the list of visible avatars (avatar 1 in the case of Figure 4.1). The resulting list is passed to the next step, which is view the frustum culling stage. The avatars out of the viewing frustum (indicated by dashed red lines) are detected and excluded from the list (avatar 2). The third step is the back face culling stage, which is performed by comparing the orientation of the remote avatars with the local avatar’s orientation. Those avatars facing away from the local avatar are culled (avatar 3). The last step is the occlusion culling stage, which is computationally the most complex process. In this stage, those avatars that have passed the previous tests are ordered based on their distance to the local avatar. The angle of the imaginary lines (solid blue line) going through the view point and the extreme points of the avatars are then calculated (α and β) for each avatar. The view field, limited to the view frustum side planes (dashed red lines) is called the visible angle range, and the angle from α to β
Figure 4.1 Visibility culling

is called the avatars’ blocking angle range.

From the nearest avatar to the furthest avatar, an overlap test is applied to check if the blocking ranges of avatars have any overlap within the total visible angle range. Each avatar’s blocking angle range is first compared to the blocked area, if there is any overlap, then further testing needs to be done, otherwise the avatar is visible. Finally, the angle range is added to the blocked area.

In Figure 4.1, avatar 4 is completely occluded by avatar 5, because avatar 5 is closer than avatar 4 to the local avatar, this avatar is tested first and its entire angle range is added to the blocked area list. When the avatar 4 is tested, the avatar’s angle range is found to overlap with the blocked area, so further tests are needed. By performing the visible area calculation, the angle range for avatar 4 is detected as fully occluded.

Algorithm 1 explains the visibility test, where ‘Angle’ only contains the relative angle to the local avatar’s orientation (Angle = avatar’s orientation angle - local avatar’s orientation angle) and ‘AvatarsPositionSortedbyDistanceAscending’ is defined to be the output of distance based culling (DBC). According to distance based culling method, if two avatars have the same distance during the sort procedure, the avatar with the closer extreme points (left or right) is placed
Distributed Area of Interest Management

foreach Object in AvatarsPositionSortedbyDistanceAscending do
    if Object.Angle is between 90 and 270 then
        if Object.AngleRange has overlap with BlockedArea then
            Object.VisibleArea ← CalculateTheVisibleAreaOf(ObjectbehindBlockedArea)
        else
            Object.IsVisible ← true
        end
    else
        Object.IsVisible ← false
    end
    BlockedArea.add(Object.AngleRange)
end

Algorithm 1: Visibility test

first in the sorted array.

4.3.1 Conservative visible set

In this study a distributed AOI is introduced, where each client calculates its own AOI according to the algorithm presented above and maintains a list of visible avatars (the visible set), which is sent to the sources of video (which could be a server, a collection of peers, or a hybrid of both depending on the architecture of the system) as a request. Due to the distributed nature of the system, delay is inevitable and must be considered. This means that the visible set should remain valid by the time the client receives the requested video streams back. Otherwise, there may be a discrepancy between the desired view of the client’s avatar and the actual information received. The round trip delay between the client and the video sources represents the latency that has to be included in designing of the video distribution control algorithm. We refer to this delay as the network response time. As a solution, a conservative visible set (CVS) is introduced, which not only contains the visible avatars at the current point in time but also the avatars that may become visible during the interval between sending the request and receiving the response.

To construct this list, it is assumed that there is a motion vector related to each avatar, which indicates the current speed and direction of an avatar. Based on
Figure 4.2 Conservative visible set

the motion vector, the position and orientation of the avatar at the end of the network response time is predicted. The avatar at the current point in time is called *current avatar* and the avatar with the predicted position and orientation is called *predicted avatars*. The Algorithm 1 is applied not only on the current avatar but also on the predicted avatar. For instance, in Figure 4.2, avatar 5 is not visible, but the predicted position of the avatar is in the predicted view frustum of the local avatar, so the list of visible avatars should include avatar 5.

4.3.2 Conservative view field

In Figure 4.3, the current view field of an avatar is indicated by a yellow sector, the visible angle range is called $\alpha_0$ and the maximum visible distance is denoted by $b_0$. 
A conservative view field (CVF) is not only the view field at the current moment in time, but extended by all possible view fields during the network response interval. CVF is constructed based on the maximum translation velocity \( v_{\text{max}} \), maximum angular velocity \( \omega_{\text{max}} \) and network response time \( t_r \).

\[
\alpha_n = \alpha_0 + (\omega_{\text{max}} \times t_r) \times 2 \tag{4.1}
\]

\[
b_n = b_0 + (v_{\text{max}} \times t_r) \tag{4.2}
\]

\( a_n \) is the visible angle range of CVF, the formula is multiplied by two, because an avatar can either turn right or left and by this multiplication all possible future view fields are covered and \( b_n \) is the maximum visible distance of CVF.

### 4.3.3 3D area of interest management

2D and 3D refer to the dimensions of the system. In a 2D model, the positions of avatars are represented by two values \((x, z)\). The change of \( x \) and \( z \) indicates the movement of avatars along the width and length of the environment respectively. However, in the 3D model, the position of each avatar is denoted by \((x, z, t)\).
The change of $y$ indicates the movement of avatars along the height of the environment. Introduction of $y$ affects the occlusion culling mechanism dramatically. Since an avatar can be located behind another avatar but not occluded, due to having a different height value. To overcome this issue, the proposed AOI mechanism is adapted to the 3D characteristics of the environment. In this approach, the bounding volumes of current avatars and the conservative view field (CVF) planes are tested by exploiting an aligned axis bounding box (AABB) view frustum culling technique. In this method, a box with the smallest area, within which, the avatar fits, is constructed. Then the box is aligned with the world axes to speed up the computation. Note that the orientation of the box is maintained during the process. When the avatar is moved and rotated, only the size and location of the box is changed to accommodate the avatar. In order to find, if the avatar is located in the conservative view frustum of the local avatar, the intersection between this box and the view frustum is calculated. If the current avatar is detected outside of the view frustum (VF), the AABB view frustum culling will be performed on the predicted avatar. Being inside of the view frustum leads to occlusion culling examination (refer to Section 4.3.4) and being outside of the view frustum means that the avatar is invisible and remains invisible until the end of the next network response time. Existence of the current avatar inside the VF cannot confirm the visibility of avatar necessarily and just conducts the test to the occlusion culling test. However, in our approach current avatars always have priority in tests to the predicted ones and detection of an avatar as ‘visible’ in any stage of tests terminates the whole test (Algorithm 2).

4.3.4 3D occlusion culling

Proposed occlusion culling (OC) method is based on ray casting. In computer graphics, ray casting is usually used to determine pixels’ colors by emitting a ray from the camera through each pixel and checking the intersection of the rays with the objects in the scene. Due to the huge number of required intersection calculation, the ray casting methods are very complex. However, the proposed ray casting method in this research is a real-time optimised technique, which
if Avatar.current is InsideViewFrustum then
    if Avatar.current is visible based on OcclusionCullingTest then
        VisibleSet.add(Avatar.ID)
    else if Avatar.prediction is InsideViewFrustum then
        if Avatar.prediction is visible based on OcclusionCullingTest then
            VisibleSet.add(Avatar.ID)
        else
            VisibleSet.Remove(Avatar.ID)
        end
    else
        VisibleSet.Remove(Avatar.ID)
    end
else if Avatar.prediction is InsideViewFrustum then
    if Avatar.prediction is visible based on OcclusionCullingTest then
        VisibleSet.add(Avatar.ID)
    else
        VisibleSet.Remove(Avatar.ID)
    end
else
    VisibleSet.Remove(Avatar.ID)
end

Algorithm 2: 3D visibility test

employs limited number of rays to perform the occlusion culling and back face culling at the same time. In this technique, the view point is not the source of rays but the destination of them. Source of rays are attached to the video surface and can make any arbitrary shape based on the content of the video. Furthermore, each vertex can be given a weight, which indicates the priority of the ray. As an example, in Figure 4.4, the sources of rays construct a triangular shape for a talking head video. Vertex A has the highest priority (weight = 2) and Vertices B and C have the same lower priority (weight = 1). Rays are emitted toward the view point in order of their priorities. This feature provides a content-aware AOI mechanism, which can be adjusted based on the available capacity of computation. In the simplest conservative model, from four corners of each avatar’s video display, rays with same weight are emitted toward the view point (Figure 4.5), if all rays are intersected by objects, including other avatars or any static or dynamic opaque object, the avatar is occluded, otherwise if even one ray is received by the view point, avatar is visible. The nice property of this novel model is that the back face culling is inherently provided. When
an avatar is rotating away from the view point, the rays would not be received and hence avatar is detected invisible (BFC).

**4.4 Simulation results**

A simulator is implemented to evaluate the amount of video downstream bitrate saved after applying the proposed AOI algorithms. The amount of saving is achieved by measuring the total downstream bitrate of IVC before and after applying the AOI mechanism. In one iteration of the simulation, 30 avatars are distributed uniformly in a fixed size (100m x 100m) environment. Then the number of visible avatars after applying each culling technique is measured. According to the simulation results shown in Table 4.1, the standard deviation of output in comparison to the mean is small after 100 iterations, and even after increasing the number of iterations to 30000 no significant change is observed. Hence the rest of the experiments are performed based on 100 iterations.
In the following experiments, the local avatar is located at the centre of the environment and the remote avatars are distributed around it. Different type of distributions are studied in the experiments. The local avatar as well as the remote ones have random orientations, unless it is explicitly expressed otherwise.

4.4.1 Impact of avatar density on total downstream bitrate

The effect of increasing the number of IVC participants from 5 to 60 in a fixed size (100m x 100m) environment is evaluated. The avatars are distributed uniformly with random orientations. The result shows that the downstream bitrate of the local avatar increases linearly when there is no AOI strategy applied, as expected. By adding a simple distance based AOI technique, a modest bitrate saving of approximately 22% is achieved. After exploiting a view frustum culling in addition to the distance based strategy, significant additional saving of 51.56% is achieved as demonstrated in Figure 4.6. The back-face culling and occlusion culling mechanisms also enhanced the results and respectively reduced the total downstream bitrate of IVC by 86.07% and 90.61%.

Figure 4.5 Ray casting mechanism
Figure 4.6 Density versus total bit rate normalised to the bit rate of one video stream (uniform distribution)
Table 4.1 Testing the solution convergence

<table>
<thead>
<tr>
<th>Iteration</th>
<th>10</th>
<th>100</th>
<th>1000</th>
<th>5000</th>
<th>10000</th>
<th>20000</th>
<th>30000</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of avatars</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>DBC</td>
<td>24.1</td>
<td>23.5</td>
<td>23.526</td>
<td>23.4544</td>
<td>23.4752</td>
<td>23.4617</td>
<td>23.46613</td>
</tr>
<tr>
<td>VFC</td>
<td>8.3</td>
<td>7.7</td>
<td>7.925</td>
<td>7.8062</td>
<td>7.8148</td>
<td>7.8241</td>
<td>7.822733</td>
</tr>
<tr>
<td>BFC</td>
<td>4.8</td>
<td>4.03</td>
<td>3.945</td>
<td>3.9214</td>
<td>3.9203</td>
<td>3.9169</td>
<td>3.903833</td>
</tr>
<tr>
<td>OC</td>
<td>3.7</td>
<td>3.3</td>
<td>3.181</td>
<td>3.1758</td>
<td>3.1493</td>
<td>3.1373</td>
<td>3.137533</td>
</tr>
<tr>
<td>stDev DBC</td>
<td>2.071232</td>
<td>2.161018</td>
<td>2.234575</td>
<td>2.267051</td>
<td>2.248285</td>
<td>2.260118</td>
<td>2.260675</td>
</tr>
<tr>
<td>stDev VFC</td>
<td>3.195309</td>
<td>2.286919</td>
<td>2.4567</td>
<td>2.379126</td>
<td>2.415223</td>
<td>2.390975</td>
<td>2.389584</td>
</tr>
<tr>
<td>stDev BFC</td>
<td>2.181742</td>
<td>1.59659</td>
<td>1.921972</td>
<td>1.833582</td>
<td>1.85417</td>
<td>1.849647</td>
<td>1.837077</td>
</tr>
<tr>
<td>stDev OC</td>
<td>1.002497</td>
<td>1.089908</td>
<td>1.137356</td>
<td>1.125062</td>
<td>1.125439</td>
<td>1.129475</td>
<td>1.12616</td>
</tr>
</tbody>
</table>

In another experiment, all avatars are distributed normally around the local avatar with a mean of (0,0) and $\sigma^2 = 50$ m$^2$ in both dimensions. The result is similar to the case of the uniform distribution, with the differences that the distance based strategy is less effective due to the normal distribution of the avatars, and the occlusion culling is more efficient and can save up to 92.8% of total downstream bitrate, because the avatars are densely clustered around the local user and cause more occlusion (Figure 4.7).

### 4.4.2 Impact of translational velocity on total downstream bitrate

In this section, a translational velocity is introduced to the system, and the prediction of the avatars’ location after a specific delay, which is equal to the network response time, is calculated. The network response time is set to 200 ms to simulate an intercontinental/national network size. It is assumed that the avatars can move only in the direction of their orientations, while the orientation and velocity is completely random as shown in Figure 4.8. In this simulation (Figure 4.9), the total number of avatars is fixed and equal to 25, the maximum velocity ($v_{max}$) varies from 1m/s to 90m/s in steps of 1m/s and the network response time is set at 200ms. Each avatar is assigned a random velocity in the range of (0, $v_{max}$).

The simulation has also been repeated for 60 avatars with virtually indistin-
Figure 4.7 Density versus total bit rate normalised to the bit rate of one video stream (normal distribution)
Figure 4.8 Introduction of translational velocity to the system
Figure 4.9 Impact of translational velocity (25 avatars)
Figure 4.10 Impact of translational velocity (60 avatars)

guishable results (Figure 4.10).

Figure 4.11 shows the percentage increase in bit rate for each method based on the uniform distribution of the same number of avatars (60 avatars) after introducing translation velocity to the system. When no AOI management is applied, all client data is sent to all other clients, so mobility has no impact on total video downstream bitrate. The distance based method is also not very motion sensitive, since it covers a large area containing most of the avatars. However, as shown in figure 4.11, motion has a strong influence on view frustum, back face and occlusion culling because predicted location of the avatars may now be in the visible area. However, the absolute bit rates are still much lower (as seen in figure 4.10). Some fluctuations are visible in behaviour of view frustum, back face and occlusion culling techniques, which reveal that the probability of entering to the viewing frustum is the same as the probability of
exiting from it and the probability of being an occluder for a predicted avatar is as high as being an occludee in the case of occlusion culling. In other words, it is more likely that avatars with higher translational velocities change their positions in the next iterations and hence change their visibility statuses from visible and/or occluder to invisible and/or occludee or vice versa.

4.4.3 Impact of angular velocity on total downstream bitrate

In this section all simulation parameters are the same as in Section 4.4.2 but instead of translational velocity, angular velocity is introduced to the system. The result shown in Figure 4.12 is achieved by varying the angular velocity from $10^\circ/s$ to $360^\circ/s$ in increments of 10 percent. From Figure 4.12, it is clear that angular velocity has a large impact on view frustum, back face and occlusion culling and almost no impact on distance based strategy. The primary reason
of this behaviour is, the rotation of the local avatar’s viewpoint and not the changes in the direction of other avatars. In the distance based method, because the local avatar is located in the middle of the circle, its rotation has no effect on the result and the slight changes are a consequence of the rotation of other avatars. However by rotating the viewing frustum, new users are added to the visible set, which is passed to the subsequent culling stages. The same behaviour is seen in back face and occlusion culling. Clearly, back face culling is vulnerable to higher angular velocity, since it is more likely for the ‘current’ avatar or ‘predicted’ one to eventually become visible by rotating toward the local avatar.

In Figure 4.13, the movement of avatars causes an increase of up to 300% for the downstream bit rate. This increase as shown in this section and Section 4.4.2, is due to the number of avatars that we would not have received their
video streams if the conservative visible set had not been introduced. While this is a significant relative cost to pay for dealing with motion, in total, the absolute bit rates compared to the case of no AOI management is still much better. In the worst case scenario that the maximum angular velocity is applied to the avatars, the AOI mechanism still reduces the total downstream by 76.3% compared to the case that the AOI method is not utilised.

4.4.4 Analysing a realistic scenario

In practice, both translational and angular velocities are present at varying degrees and the users may be clustered together. Therefore, in this test to simulate a real scenario, it is assumed that the maximum possible translational and angular velocities are respectively 15 m/s and 180°/s and corresponding avatar velocities are uniformly distributed in these ranges. A network response time of 200 ms is added to the system and in different tests 25 avatars are clustered.
around different numbers of centre points, facing their respective centre points with an offset of 15°. The local avatar is placed as one of these centre points. In a 100m x 100m environment a maximum distance of 15 meters to the centre point is used. The demonstrated results are obtained by increasing the number of centre point from 1 to 6. As it is shown in Figure 4.14, when there is one centre point, because the local avatar is located at the centre point and all other avatars are looking toward the point, back face culling is ineffective but by increasing the number of centre points the result becomes similar to the uniform distribution. In all cases, occlusion culling is highly effective, due to the high density of avatars around the local avatar. It is also clear that by using the AOI algorithms introduced in this chapter, the required downstream capacity of IVC clients is significantly reduced by up to 76.44%.
4.5 Conclusions

In this chapter, a real-time AOI strategy for downstream bitrate reduction is implemented and the impact of different factors, including distribution, density and speed is analysed. The result indicates that the combination of proposed distance based, view frustum, back face and occlusion culling techniques can decrease total downstream bitrate up to 90%. Due to the feedback delay in the system, a conservative visible set needs to be created and consequently, a prediction of the position and orientation of the avatars must be calculated which is dependent on their respective translational and angular velocity. The results reveal that the angular velocity has a greater effect on video downstream bitrate than translational velocity. However a significant total downstream bitrate saving is still achieved even at high avatar translational and angular velocity. Finally, a more realistic scenario based on a clustered model is explored and the results demonstrate the applicability of the developed AOI strategy in an immersive multimedia system. By exploiting the AOI mechanism, transmission of invisible avatars’ videos is avoided. However, the same share of downstream bitrate is allocated to the visible avatars, while their avatars may not appear the same size on the screen, due to their relative virtual positions and orientations. In Chapter 5, the hypothesis, that the avatars with smaller projection sizes on the screen require lower video resolution and hence less network capacity is analysed.
Chapter 5

Evaluation of User Sensitivity to Video Quality Differentiation

5.1 Introduction

The area of interest (AOI) management is discussed in Chapter 4. Utilisation of this method in IVC reduces the required network capacity by preventing transmission of invisible avatars’ video. In addition to AOI, we believe there is an opportunity to reduce the video downstream bitrate through differentiating the video quality of visible avatars. Hence, in this chapter, a video quality differentiation (VQD) mechanism is introduced to further improve the scalability of IVC. The enhancement is achieved by adjusting the video quality sent to each client based on the virtual positioning and orientation of that client’s avatar in the 3D environment. To find the perceptual threshold of the participant for detecting degradation in video quality, a subjective video quality assessment in the context of a representative 3D IVC is conducted. The study included 12 different questions analysing the impact of the avatars’ virtual positions and orientations on the perception of spatial and temporal degradation of video quality. Since laboratory quality studies are time-consuming, expensive and limited in number of participants, a web-based mean opinion score (MOS) video quality assessment is used. Since it is not possible to supervise the participants during this process, a method is developed to detect and discard statistically invalid scores. A total of 233 individuals participated in the study; however, each subject was
free to skip or re-do each question, with the exception of the final two.

This chapter is structured as follows: in Section 5.2 the design procedure of the subjective study is discussed; in Section 5.3 the process of subjective scores and the proposed models are introduced; Section 5.4 presents simulation results and analysis; and finally Section 5.5 summarises the conclusions of the chapter.

5.2 Overview of Subjective Study

A survey 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. In the first three questions the avatars are located at 9, 6 and 3 meters distance from the viewpoint respectively and the spatial resolution of the video attached to the target avatar is degraded. In question three to six, the same virtual distances are applied to the avatars, while the temporal resolution of the target videos is dropped. From question seven to ten, the effect of virtual orientation on the perceived video quality is studied. In question seven and eight, avatars are rotated about their local y axes by 30 and 60 degrees respectively, while the spatial resolution of the target video is decreased gradually. In question nine and ten, avatars have the same orientations, however, the joint impact of orientation and reduction of temporal resolution is analysed. The last two questions investigate the impact of the viewer’s focal point on change of spatial quality. One frame of each category in terms of virtual situation is demonstrated in Figure 5.1.

A degradation category rating (DCR) scheme, also known as double stimulus impairment scale (DSIS), was adopted for this study [84]. 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
Figure 5.1 One frame from each of the reference and target avatar used in the study
the study is as follows:

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

### 5.2.1 Reference sequences

To obtain reference sequences, 14 ‘talking head’ videos in common intermediate format (CIF) resolution (352×288) are captured (Figure 5.1). Since the user study is conducted for video conferencing in the context of IVC and projected size of avatars and hence required video resolution is mainly dependant to the virtual position of avatar and IVC’s resolution, the CIF quality appears to be an appropriate size for the study. 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 (FPS).

### 5.2.2 Target sequences

120 target sequences are created by degrading 6 second (120 frame) slices of the 12 reference sequences (Figure 5.2). 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. Let

---

1 According to the average chosen size of IVC window by the users (640×480 to 1152×864) and projection sizes of the video surface on the screen, when the avatar is completely visible in the viewer’s view frustum, the CIF size is selected. This topic is extensively analysed in Chapter 6
Table 5.1 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>

$Res_{(j)(k)}$ denote the spatial or temporal resolution of slice $k$ of video $j$. Then:

$$Res_{(j)(k)} = Res_{(j)(k-1)} - \lfloor 0.2Res_{(j)(k-1)} \rfloor$$  \hspace{1cm} (5.1)

The spatial and temporal resolution step-sizes are given in Table 5.1.

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 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 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.

### 5.2.3 Subjective test design

In order 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
Evaluation of User Sensitivity to Video Quality Differentiation

Figures 5.2 Subjective assessment setup

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.

5.2.4 Presentation of test questions

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 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 5.2.5. The IVC was recorded at 25 frames per second to avoid any frame drop. To make the web-based study cross-platform and accessible to majority of the users, the recorded process 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 and then released for playback. The remaining area around the video was white, with a progress bar and the ITU-R ACR
scale (with modified wordings) displayed at 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 Figure 5.3. 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.

To prevent internet bots from accessing the study, captcha protection was used. Furthermore, an email verification system was employed to prevent the participation of users with invalid or fake identities.
5.2.5 Detailed description of the subjective study

In this research, IVC is a game-like 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 (see Chapter 3). The major bottleneck for such a system is the network transmission capacity needed to support the service. In the last chapter, a range of real-time strategies were proposed to minimise the downstream bitrate consumption and hence improve the system’s scalability (Chapter 4). The main goal of this chapter 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.

5.2.5.1 Virtual distance vs. spatial resolution

In Question 1, reference and target avatars are located 9 meters (in the virtual environment) away from the viewpoint (Figure 5.1 (A)). The target sequence containing ten slices with ten different spatial resolutions (100%-13.6% CIF) (Table 5.1) 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 Hz is paired and 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. Note that the 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 6 and 3 meters from the viewpoint, and the same process of degradation is performed (Figure 5.1 (B) and (C)).

5.2.5.2 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 dropped from 20 fps to 1 fps as described in
Section 5.2.2. As for the first three questions (discussed in Section 5.2.5.1), 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.

5.2.5.3 Virtual orientation vs. spatial resolution

In Questions 1 to 6, the avatars were facing directly toward the viewpoint. Note that, 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 about their y axes, respectively (Figure 5.1 (D) and (E)).

The same spatial degradation effect as described in section 5.2.5.1 was applied to the target avatar and the scores were recorded.

5.2.5.4 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 (Table 5.1) and the impact of virtual orientation on the users’ sensitivity to detecting the temporal resolution degradation was investigated.

5.2.5.5 Viewer’s focal point vs. spatial resolution

In questions 11 and 12, although two avatars were presented to the subject, neither of them were labelled as the reference avatar. Both avatars had a red dot on the top and were labelled as a “target” avatar. One of the avatars was located closer to the viewpoint 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 reso-
lution (CIF) was describing the instructions for the question to the subject, the spatial resolution of the more distant avatar was being degraded gradually (Figure 5.1 (F)). 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 changed slightly in 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, the video of different people with diverse hand gestures were captured and displayed.

5.3 Subjective scoring methodology

Since the conducted study is of the degradation category rating 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:

$$d_{ijk} = 5 - s_{ijk} \quad k = \{1, 2, 3, ..., 10\} \quad (5.2)$$
Then z-scores\(^2\) per slice are calculated:

\[
\mu_{ik} = \frac{1}{N_{ik}} \sum_{k=1}^{N_{ik}} d_{ijk}
\]

\[(5.3)\]

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

\[(5.4)\]

\[
z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\delta_{ik}}
\]

\[(5.5)\]

where \(N_{ik}\) is the number of slices scored by subject \(i\) across all experiments. 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 applied to the results to reject unreliable subjects [85]. 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 the scores assigned by him/her 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 the scores assigned by him/her fall outside the range of 4.47 standard deviations from the mean scores. Since in our study a question can be answered multiple times, first the subjects with more than one submitted scores to a slice are found. Then, one of the submitted scores from that particular subject which was closer to the mean value was recognised as the answer and the rest were removed. Finally, the unreliable subjects were detected and eliminated, resulting in the removal of 20 out of 233 subjects.

\(^2\)Standard scores also known as z-scores are a statistical measurement which represent the relationship of standard deviation to the mean. A positive z-score indicates a datum above the mean, while a negative standard score represents a datum below the mean.
5.3.1 Data processing

Figures 5.4 and 5.5 demonstrate the results of the subjective study for the first 6 questions. In order to extend the results achieved in Chapter 4, the developed simulator is exploited in this chapter as well. However, the user study is conducted in the actual IVC system. To make the ratio of avatars to the environment consistent through both entities, the values are scaled 5 to 1. Hence, every 5 meter in the simulator world is equal to 1 meter in the 3D world of IVC. As the result of this scaling, the 45, 30 and 15 meters shown in the figures are exactly the same as 9, 6 and 3 meters in the conducted user study. As expected, regardless of the virtual distance, recorded MOSs reduce as the spatial or temporal resolutions decreases. However, the closer the avatar is located to the camera, the more quality degradation is perceivable. For example, a degradation of 48.6% (spatial resolution of $181 \times 147$ pixels) is acceptable (MOS above 3.5) for a more distant avatar but will be unacceptable (MOS below 2) for the closer avatar.

As shown in Figures 5.4 and 5.5, the impact of virtual distance is more pronounced on spatial resolution than temporal; nevertheless, the perception of frame rate reduction is also affected when the avatar is located at the furthest virtual position (Figure 5.5).

Figure 5.6 and 5.7 present the impact of orientations of avatars 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 that avatars are oriented directly toward the camera. As shown in Figure 5.6 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 (Figure 5.7). However, the skewness of scores are smaller for the higher angular states, which means more subjects have detected the degradation in the lower temporal resolutions, when the avatar is rotated away (Figure 5.8). The behaviour of subjects that they could not detect the spatial resolution reduction, when the avatar’s virtual distance was increasing and they
Spatial distance and resolution impact

**Figure 5.4** Impact of virtual distance on spatial resolution

Spatial distance and frame rate impact

**Figure 5.5** Impact of virtual distance on temporal resolution
could detect the degradation when avatar's orientation was increasing suggest a very close link between the projection size and shape of the video surface and the degradation method that should be employed to make it unnoticeable. It can be clearly observed that the size of the video surface is uniformly decreased 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 analysed in Chapter 6 and a degradation method is proposed that not only performs based on uniform size reduction of the video surface, but also adapts to non-uniform situation such as rotated video surface.

The influence of focal point is studied in the last two questions, where a video with audio component is assigned to an avatar and the spatial video resolution of the other avatar is degraded gradually. In the first test the further avatar's video is affected, while the avatar with audio component in the closer one trying to grab the attention of the subjects by describing the test to them. The same procedure is applied to the last question, while the roles of avatars are swapped.
Evaluation of User Sensitivity to Video Quality Differentiation

Figure 5.7 Impact of orientation on temporal resolution

Figure 5.8 Skewness of “noticeable degradation” PDF for Questions 3, 9 and 10
and the locations of the avatars are slightly changed.

As demonstrated in Figure 5.9, 41 percent of the subjects participated in question 11 did not notice any change in the quality of video and 27 percent perceived some changes. However, 34 percent of these subject detected wrong or unrelated quality issues, such as change of quality in video of the wrong avatar, lip sync and audio issues, jerkiness and frame rate drop. Rest of the subjects didn’t submit any score.

In question 12, due to the location of avatars, 49 percent of the participants were able to detect the video degradation. However, 20 percent of the answers reported invalid issues like the previous question. 37 percent of the subject did not perceive any quality change and 14 percent of the subjects did not submit any score. We can conclude, therefore, that the focus of attention can play a significant role in selective degradation of video streams.

5.3.2 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 (Figure 5.10 - green windows), even though the size of the user’s viewing window (IVC window, i.e. the image seen by his or her virtual camera in the 3D world) is fixed (Figure 5.10 - blue windows). This reduction in video size and resolution is purely due to geometry and one of the aims of the user study was to assess if the user perception matches the level of resolution decrease as predicted by this model or there are other factors involved.

In order to find the perceptual relationship between the virtual distance of 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
calculated using equation 5.6. As shown in Figure 5.11 the projected size is proportional to unprojected coordinates multiplied by the virtual distance divided by the near plane distance. Then a prediction model is obtained to predict the spatial resolution in pixels based on the virtual distance in meters. Note that based on the avatars and environment sizes in IVC with respect to the simulator, the values are scaled to make the simulator data consistent with the data in the actual environment\(^3\).

\[
\begin{align*}
    y_s &= \frac{d}{z} y \\
    x_s &= \frac{d}{z} x
\end{align*}
\]  

(5.6)

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. A three-parameter exponential function\(^4\) is then fitted to the subjective quality scores.

---

\(^3\)Since, the model is used in the developed simulator utilised in this research, the values are scaled up to be consistent with the simulator scales. Nevertheless, this conversion has no impact on the achieved results.

\(^4\)Different functions are tried for this purpose. However, the least error between the vector of subjective study and the prediction model with the least complexity is achieved by three-parameter exponential function.
Let $s$ represent the spatial resolution that the perceptual model predicts for an avatar at virtual distance $\beta$.

$$s = \alpha_1 + \alpha_2 e^{-\alpha_3 \beta}$$ (5.7)

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 5.12). The achieved model not only is capable of adjusting spatial resolution based on virtual distance, but also minimises the required spatial resolution according to perceived quality by human visual system, so that the degradation is not noticeable to the clients.

\textsuperscript{5}This function is utilised to get a vector of estimated coefficients for the nonlinear regression of the responses of the prediction model based on the submitted scores.
5.3.3 Bit-rate model

Many studies have addressed and modelled the impact of spatial, temporal and amplitude resolutions on perceptual quality [86, 87]. However, some of the proposed models are computationally expensive, due to the number of features and parameters used [88]. In [89], a bit rate model as a function of quantisation parameter 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 Zhan et al. [90]. However, the spatial resolution is ignored in the proposed models. In this section, an analytical model for video bit rate in terms of spatial and temporal resolutions is proposed. The model proposed in [90] is extended to account for spatial resolution. As will be shown later, the model accurately predicts the bit rate from spatial and temporal resolution of the videos.

In this work, the focus is on spatial and temporal resolutions, while amplitude resolution will be studied in future work. Toward this goal, the bit rate model is considered as a function of spatial and temporal resolutions and hence the bit rate $R(s, t)$ is written as:

$$R(s, t) = R_{\text{max}} R_s(s, t_{\text{max}}) R_t(t, s)$$  \hspace{1cm} (5.8)

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

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

$$R_s(s, t_{\text{max}}) = \frac{R(s, t_{\text{max}})}{R(s_{\text{max}}, t_{\text{max}})}$$  \hspace{1cm} (5.9)

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

\[ R_t(t, s) = \frac{R(s, t)}{R(s, t_{\text{max}})} \quad (5.10) \]

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 5.2.2 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 5.1. The bit rate for each sequence is calculated\(^6\) 5.13. Then 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 Figure 5.14.

The curves achieved by different frame rates overlap with each other and can be characterised by a single curve. The behaviour demonstrated in Figure 5.14 suggests that the impacts 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 \(t\).

It is discussed in [90] 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 also shown to be a power function, explained as follows:

\[ R_t(t) = \left( \frac{t}{t_{\text{max}}} \right)^b \quad b \leq 1 \quad (5.11) \]

\(^6\)The sequence is neither in the raw data format such that the bit rate could be calculated simply via multiplying the size of each frame by the frame rate, nor completely coded via a video codec to introduce the codec’s specifications and consider the codec’s behaviour in the modelling process. Since, the data is only transmitted locally (from the decoder to IVC), the frames are partially decoded in a way that they are still applicable inside the 3D environment and computationally less expensive to handle inside the IVC system. Moreover, the transmitted bitrate is reduced to some extent and the frames can be utilised as textures inside IVC (a jpeg like frames are achieved in this stage). Hence, the bitrate of the system needs to be modelled.
Experimental data also confirms 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 of the suggested function to model the system based on spatial resolution is:

$$R_s(s) = \left( \frac{s}{s_{\text{max}}} \right)^d \quad d \leq 1$$  \hspace{1cm} (5.12)

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 before 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 [90]. The parameter $d$ had the value of 0.6312. Other functions such as logarithmic and inverse falling exponential were also tried to fit the curve, but the power function yields the minimum
Combining Equations 5.11 and 5.12, the following overall rate model is proposed:

$$R(s, t) = R_{\text{max}} \left( \frac{s}{s_{\text{max}}} \right)^d \left( \frac{t}{t_{\text{max}}} \right)^b$$  \hspace{1cm} (5.13)

where $s_{\text{max}}$ and $t_{\text{max}}$ are the maximum spatial and temporal resolutions respectively and should be set based on the required applications. $R_{\text{max}}$ 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. 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 resolu-
Figure 5.15 Bit rate vs. temporal resolution

...was calculated while the spatial resolution was fixed at CIF quality. Then, using the proposed model the bit rates were predicted; the results are shown in Figure 5.15. 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.

5.4 Simulations

In Chapter 4, a range of strategies for real-time area of interest (AOI) management in IVC were proposed to minimise the required video transmission capacity. A simulator was implemented to analyse the amount of downstream
Figure 5.16 Bit rate vs. spatial resolution

bitrate 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 5.13 and 5.7 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 × 100 m) immersive environment and the simulations are performed with 100 iterations in which the clients are distributed uniformly with random orientations, unless otherwise specified.
5.4.1 Impact of density on downstream bitrate

The impact of density on downstream bitrate is evaluated by increasing the number of clients from 5 to 60 in the IVC. In Chapter 4, it was shown that by exploiting the AOI methods, a total downstream bitrate saving of 90.61% can be achieved when 60 randomly distributed clients are present in the environment. After adding the VQD mechanism, not only any unnecessary transmission of video is stopped but also the quality of video is degraded in a way which is not perceptible to the viewer, and the total downstream bitrate saving reaches 96.85%. This is shown in Figure 5.17.

When the spatial distribution of avatars is changed to a two-dimensional normal distribution centred around the local client with and $\delta^2 = 50 \text{ m}^2$ in both dimensions, due to occlusion culling, the AOI strategies are more effective and
can save up 92.8% of downstream bitrate, while the VQD is less efficient. This is because all avatars are densely clustered around the local client and require the highest video quality. Nevertheless, after applying the VQD method the saving increases to 96.08%. This is shown in Figure 5.18.

5.4.2 Impact of translational 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 Chapter 4 was utilised to predict the client’s next position after the 200 ms simulated network response time.

The result shows that VQD can increase the saving by an average of 67.13% when 60 clients are present in the IVC. Results are shown in Figure 5.19.
Figure 5.19 Impact of translation velocity on bit rate
5.4.3 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 response time is set to 200 ms.

The behaviour of the system as studied extensively in Chapter 4 showed that an increase of 300% in required network capacity is possible when angular motion is introduced. The AOI mechanism saves significant amount of downstream bitrate. However, after applying the VQD to the system, the savings are carried on and increased to an average downstream bitrate saving of 93.54%. Results are shown in Figure 5.20.
5.4.4 Analysing a realistic scenario

In order to simulate a more realistic scenario, both translational and angular velocities of the users are varied in this experiment. The avatars are clustered around a few centre points, where 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 are 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 Figure 5.21, occlusion culling is highly effective due to the high density of clients around the local client. Nevertheless, VQD can still further improve the downstream bitrate 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.

5.4.5 Analysing 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 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 hypothetical scenario, the AOI mechanism is ineffective and does not reduce the required network capacity requirements for the lecturer’s client. Such a scenario is shown in Figure 5.22

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 it is expected, the AOI methods are entirely ineffective.
Figure 5.21 Impact of clustered distribution on bit rate
Figure 5.22 Lecture theatre distribution
However, after exploiting the VQD strategy, a significant average downstream bitrate saving of 74.60% is achieved. This result is shown in Figure 5.23.

5.5 Conclusion

In this chapter, a subjective study was presented, which aims 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 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 is 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 enables us to spatially and temporally degrade the video quality so as to have 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 evaluated. 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 downstream bitrate saving can be achieved in all scenarios even when the AOI mechanism is completely ineffective. However, in this method the spatial resolutions of the videos are degraded uniformly, which does not perform well in virtual situations, where the projection of the video surface is distorted (e.g. rotated video surface). In the next chapter a method is proposed that applies different level of degradation to different spatial regions of the frame based on the projected shape and size of the video surface.
Chapter 6

Perceptual Pruning

6.1 Introduction

An immersive video conferencing system delivers video data to 3D virtual environments in order to provide a novel and engaging interactive experience. The sense of immersion is created by attaching videos to avatars in a 3D environment. This process is performed by breaking down videos into individual frames and applying each frame as a texture onto the video surface. Then, the surface with the frame on top is projected on the screen. Mapping from the texture space to screen space is divided into two phases. Initially, the texture is mapped from texture space to object space and then the modelling and view transformations will map object space to screen space usually by exploiting a perspective projection. A pixel in texture is called a texel. Due to the perspective projection, multiple texels may be projected onto a single pixel or, conversely, a single texel may be shown as many pixels on the screen. If the ratio of texel to pixel is less than one, then this process is called texture magnification; otherwise it is called texture minification. Different texture filtering mechanisms have been developed to compute the contribution of texels to each of the screen pixels. Texture mapping and filtering are out of the scope of this research. However, this chapter aims to take advantage of texture minification and magnification to minimise the required downstream capacity for transmission of video data with negligible perceptual impact.
6.1.1 Overview of the proposed method

The first method to reduce the required network capacity was introduced in Chapter 4. Using the area of interest management technique, the system dynamically evaluates which of the avatars are within the visual range of the viewer. Consequently, only those videos that are relevant to the AOI are transmitted to the client, which results in a significant reduction in overall consumption of network capacity.

It was then shown in Chapter 5 that further reduction of downstream bit rate could be achieved by exploiting video quality differentiation, which adjusts the bit rate of individual videos within the same AOI based on the 3D situation of the avatars. However, both perceptual quality and bit rate are not optimised in the situation that the projections of avatars are distorted. It is therefore proposed to develop techniques which allow individual video streams to be pruned before transmission to the client based on their exact projection sizes. The pruning can take place at the source of each video stream in a P2P model or in the server for a server-based IVC. Note, however, that different participants will have different perspectives. Therefore a particular video stream may be required at many different quality levels depending on the number and relative position of other participants who are looking at that particular avatar at this time. Hence, unlike a point-to-point video telephony scenario, it is not possible for the source to simply adjust its video coding parameters based on the receiver’s requirement.

Moreover, the exiting models for video quality differentiation designed for multicast video distributions either exploit hierarchical video coding (HVC) or multiple description coding (MDC), which are not suitable to be utilised in the IVC system. As degradation of video quality in HVC and MDC is uniform and implemented based on visual characteristics of human eyes and not 3D projection of the video content (refer to Section 2.3).

This Chapter is structured as follows: in Section 6.2 to 6.7 the transformation process is discussed; Section 6.8 analyses the blocks’ behaviours in differ-
ent 3D situations; in Section 6.9 a proposed degradation procedure and DCT
downsampling method are introduced; Section 6.11 presents simulation results
demonstrating the efficacy of the proposed methods; and finally Section 6.12
summarises the conclusions of the Chapter.

6.2 Transformation of macroblocks

While a texture is mapped to the screen, the scaling ratio of texel to pixel
in different areas of a texture is usually different. However, calculation of the
projection ratio of each texel for the purpose of this research is not required
and causes unnecessary computational cost. Thus, the texture is divided into
spatial regions and the ratio of each region is calculated. On the other hand,
in IVC systems, textures are actually video frames and each frame is already
divided into spatial regions known as macroblocks.

Modern video codecs, like H.264/AVC use the notion of macroblock partition to
refer to the group of pixels in a block that share a common prediction. Basically
encoders use search methods to determine suitable motion vectors and compare
the rate-distortion score for possible macroblock partitioning e.g. 16×16, 16×8,
8×8, 4×4 blocks, such that the video coding performance is optimised.

In this research, the transformation from texture space to screen space is devel-
oped in a way that the texture can be partitioned to any number of arbitrary
sized regions. However, to minimise computation and enhance performance,
the transformation is adapted to use video macroblocks. Thus, the term block
refers to video macroblocks, unless explicitly stated otherwise.

6.2.1 Introduction to transformations

In traditional video conferences, the video is always received by the viewer
in a perfect rectangular shape. However, in IVC systems, as a result of the
3D presentation of avatars and perspective projection, the video surface may
be distorted. This distortion is usually not uniform and some macroblocks
in the frame may be compressed (texture minification) while others may be expanded (texture magnification) (Figure 6.1). The main goal of this chapter is to calculate the size of each block’s projection on the server or transmitting peer’s side to discard unnecessary information. Therefore, in the proposed method, rather than upsampling or downsampling the whole frame uniformly, the size of macroblocks’ projections are calculated and the resolution of each block is adjusted independently.

We use coordinate frames and assume that we have an origin $\Phi$ and three mutually perpendicular axes in the direction of $i$, $j$ and $k$. The position of each avatar is defined by point $P$ which represents the position of the centre point of the graphical representation of the avatar in the 3D environment from the origin. Point $P$ in the frame is given by $P = \Phi + P_x i + P_y j + P_z k$ and so the homogeneous presentation of $P$ is:

$$
\tilde{P} = \begin{pmatrix}
    P_x \\
    P_y \\
    P_z \\
    1
\end{pmatrix}
$$

\textbf{Figure 6.1} Projected size of the macroblocks after 60 degree rotation about local $y$ axis.
Although avatars in IVC systems are free to move and rotate, it is assumed that they are only allowed to rotate about their local $y$ axis (the axis passing through the centre of the avatar and parallel to the global $y$ axis). However, the mathematical model developed in Section 6.6 can easily be extended for a full range of motions.

Rotation around the $y$ axis Euler angle ($y$ roll) is mostly considered in this Chapter due to the aforementioned restrictions in avatar movement. The angle is denoted by $\beta$ and positive values of $\beta$ represent a clockwise (CW) rotation about the local $y$ axis as one looks inward from a point on the axis toward the avatar (Figure 6.2).
6.3 3D Affine Transformations

Affine transformations are the cornerstone of computer graphics. Any transformation can be produced by four elementary transformations: translation, scaling, rotation and shear. A succession of affine transformations can easily be combined into a single overall affine transformation.

Suppose \( T() \) is an affine transformation that transforms point \( \tilde{P} \) to point \( \tilde{Q} \). Then \( T() \) is represented by a \( 4 \times 4 \) matrix called \( \tilde{M} \).

\[
\tilde{M} = \begin{pmatrix}
m_{11} & m_{12} & m_{13} & m_{14} \\
m_{21} & m_{22} & m_{23} & m_{24} \\
m_{31} & m_{32} & m_{33} & m_{34} \\
0 & 0 & 0 & 1 \\
\end{pmatrix}
\] (6.2)

\( \tilde{Q} \) can be computed by multiplying \( \tilde{P} \) by matrix \( \tilde{M} \):

\[
\begin{pmatrix}
Q_x \\
Q_y \\
Q_z \\
1
\end{pmatrix} = \tilde{M} \begin{pmatrix}
P_x \\
P_y \\
P_z \\
1
\end{pmatrix}
\] (6.3)

6.3.1 The elementary 3D transformations

In this section, the elementary transformations are introduced individually.
6.3.1.1 Translation

The translation of an avatar in the 3D environment into a different position is achieved by the matrix:

\[
\begin{pmatrix}
1 & 0 & 0 & m_{14} \\
0 & 1 & 0 & m_{24} \\
0 & 0 & 1 & m_{34} \\
0 & 0 & 0 & 1
\end{pmatrix}
\] (6.4)

In other words, point \( \tilde{P} \) is simply translated in \( \tilde{Q} \) by the vector \( m = (m_{14}, m_{24}, m_{34}) \).

6.3.1.2 Scaling

Scaling changes the size of an object and involves three scale factors \( s_x \), \( s_y \) and \( s_z \) for the \( x \), \( y \) and \( z \) coordinates respectively:

\[
\begin{pmatrix}
s_x & 0 & 0 & 0 \\
0 & s_y & 0 & 0 \\
0 & 0 & s_z & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\] (6.5)

Note that the scaling is centred on the origin.

6.3.1.3 Shearing

A shear operation performed along the \( x \) axis does not affect the \( y \) and \( z \) components of that point, whereas the \( x \) component is translated by an amount that increases linearly with \( y \) and/or \( z \). The matrix for the simplest shear is
the identity matrix with one of the zeros replaced with a value $f$:

$$
\begin{bmatrix}
1 & 0 & 0 & 0 \\
f & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
$$

(6.6)

6.3.1.4 Rotation

In order to simplify a rotation in three dimensions, the rotations are decomposed into the simpler operations around each of the coordinate axes. The matrices that produce the rotation about $x$, $y$ and $z$ axis are presented below. In each matrix, the rotation is through angle $\beta$ about the given axis. The positive angle is defined by the looking inward convention (Figure 6.2).

The following matrices perform transformations that rotate point $P$ through angle $\beta$ about an axis. The notation $R_x()$, $R_y()$ and $R_z()$ are used to denote rotations about $x,y$ and $z$ axis respectively. The angle parameter is given in radians and $c$ stands for $\cos(\beta)$ and $s$ for $\sin(\beta)$.

$$
R_x(\beta) =
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & c & -s & 0 \\
0 & s & c & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
$$

(6.7)

$$
R_y(\beta) =
\begin{bmatrix}
c & 0 & s & 0 \\
0 & 1 & 0 & 0 \\
-s & 0 & c & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
$$

(6.8)
\[ R_z(\beta) = \begin{pmatrix} c & -s & 0 & 0 \\ s & c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \] (6.9)

### 6.4 Composition of block transformations

To achieve the aim of per-block bit rate minimisation, it is necessary to calculate the size (width and height) of each block’s projection. This is a straightforward calculation if the position of the projected block’s corner points on the screen are known - the block width and height are given by the difference between the minimum and maximum \(x\) and \(y\) coordinates of each of four the corners in the projected block.

Object coordinates (coordinates relative to the centre of the avatar) of the blocks should firstly be transformed into world coordinates. The *model transformation* is applied to each block to transform it into world coordinates.

In the process of composing the overall matrix, multiple transformations are involved. Suppose that two transformations are represented by matrices \(\tilde{M}_1\) and \(\tilde{M}_2\). Thus \(P\) is first transformed to \(\tilde{M}_1 P\) and then transformed to \(\tilde{M}_2(\tilde{M}_1 P)\). By associativity, this is just \((\tilde{M}_2 \tilde{M}_1)P\). Thus the overall transformation can be represented by a single matrix:

\[ \tilde{M} = \tilde{M}_2 \tilde{M}_1 \] (6.10)

When homogeneous matrices are used, composing complex affine transformation is achieved by a simple matrix multiplication by the product of the individual transformation matrices. The order of the matrices is reverse order to the order that transformations are applied.
Figure 6.3 Two projected blocks, illustrating the need to consider all four corners of the projected block in order to correctly calculate the height of the projected block.
6.5 Model transformation

Let \( w_i \) and \( h_i \) denote the width and height of the texture image (frame) respectively. Since the texture is defined in pixels, the position of top left corner of the texture can be denoted by \((0, 0)\) and the bottom right corner of the texture by \((w_i, h_i)\). Suppose that the spatial states of Macroblock \( m \) are defined by \( x_m, y_m, w_m \) and \( h_m \), where \( x_m \) and \( y_m \) represent the spatial position of top left corner of the block and \( w_m \) and \( h_m \) are respectively the width and height of the block in pixels.

If the size of the avatar based on world’s units is represented by \( A \), then \( A_x, A_y \) and \( A_z \) are width, height and thickness of the avatar respectively. In order to transform the block into world coordinates, the relative position of each block with respect to the centre of avatar is calculated, then the avatar is translated to its position in the virtual environment. Thus, calculating the centre position of Macroblock \( m \), when the avatar is located at \( x, y \) and \( z \) with angle of \( \beta \) can be achieved by the following elementary transformations:

1. Translate point \( P \) through vector \( v_2 = \left( \frac{w_i}{w_i} - x_m - \frac{w_m}{w_i} A_x, \frac{h_i}{h_i} + y_m + \frac{h_m}{h_i} A_y, -\frac{A_z}{2} \right) \);
2. Rotate about the \( y \) axis through angle \( \beta \);
3. Translate back point \( P \) through vector \( v_3 = (x, y, z) \);

Create a matrix for each elementary transformation and multiply them to produce an overall matrix:

\[
\begin{pmatrix}
1 & 0 & 0 & x \\
0 & 1 & 0 & y \\
0 & 0 & 1 & z \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
c & 0 & s & 0 \\
0 & 1 & 0 & 0 \\
-s & 0 & c & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\frac{w_i}{w_i} - x_m - \frac{w_m}{w_i} A_x \\
\frac{h_i}{h_i} + y_m + \frac{h_m}{h_i} A_y \\
-A_z & 0 & 0 & 1
\end{pmatrix}
\]

(6.11)
This results in the overall matrix:

\[
\begin{pmatrix}
    x - \frac{A_z s}{2} + A_x c \frac{w_m - w_i}{w_i} - \frac{w_i + x_m}{w_i} \\
    y - A_y \frac{h_m - h_i}{h_i} + \frac{y_m}{h_i} \\
    z - \frac{A_z c}{2} - A_x s \frac{w_m - w_i}{w_i} + \frac{x_m}{w_i} \\
    1
\end{pmatrix}
\]  

(6.12)

Thus \( X, Y \) and \( Z \) as the position of the centre of the Macroblock \( m \) can be achieved by Equation 6.13 to 6.15.

\[
X = x - \frac{A_z s}{2} + A_x c \frac{w_m - w_i}{w_i} - \frac{x_m}{w_i} \quad (6.13)
\]

\[
Y = y - A_y \frac{h_m - h_i}{h_i} + \frac{y_m}{h_i} \quad (6.14)
\]

\[
Z = z - \frac{A_z c}{2} - A_x s \frac{w_m - w_i}{w_i} + \frac{x_m}{w_i} \quad (6.15)
\]

By replacing \( \frac{w_m}{2} \) and \( \frac{h_m}{2} \) with zero, the world position of top left corner of the block can be calculated. Similarly, the position of the bottom right corner of the block in world coordinates can be achieved by replacing \( \frac{w_m}{2} \) and \( \frac{h_m}{2} \) with \( w_m \) and \( h_m \) respectively. The corner positions are required for calculating the width and height of the block’s projection on the screen.

### 6.6 View transformation

Since the IVC system in this research is a distributed system and each avatar views the world from its own perspective, the blocks’ position should be calculated from each avatar’s viewpoint independently. Hence, the world coordinates should be transformed to *eye coordinates* for the corresponding viewer. As explained before, world coordinates are the coordinates of the object relative to the world origin. The *view transformation* is applied to each block in the world to transform it into eye coordinates. Eye coordinates are the coordinates of the blocks relative to the camera, which indicates the local client’s view in IVC.
Furthermore, the local avatar might not be located at the origin and aligned with z axis. Hence, the whole coordinate system should be moved to the local avatar's position and rotated to be aligned with the viewer’s angle. If the position of the local avatar is represented by \(x_0, y_0\) and \(z_0\) and its orientation with \(\alpha_0, \beta_0\) and \(\gamma_0\) then the overall view transformation can be composed by the following elementary transformations:

1. Rotate about the z axis through angle \(-\gamma_0\);
2. Rotate about the y axis through angle \(-\beta_0\);
3. Rotate about the x axis through angle \(-\alpha_0\);
4. Translate back point \(p\) through vector \(v_4 = (-x_0, -y_0, -z_0)\);

\[
V = R_v(-\gamma_0)R_v(-\beta_0)R_v(-\alpha_0)
\begin{pmatrix}
1 & 0 & 0 & -x_0 \\
0 & 1 & 0 & -y_0 \\
0 & 0 & 1 & -z_0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

(6.16)

where \(R_v(-\gamma_0)\), \(R_v(-\beta_0)\) and \(R_v(-\alpha_0)\) represent the rotation matrices about \(x\), \(y\) and \(z\) axes and can be computed as follows:

\[
R_v(-\gamma_0) = \begin{pmatrix}
\cos(\gamma_0) & \sin(\gamma_0) & 0 & 0 \\
-\sin(\gamma_0) & \cos(\gamma_0) & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

(6.17)

\[
R_v(-\beta_0) = \begin{pmatrix}
\cos(\beta_0) & 0 & -\sin(\beta_0) & 0 \\
0 & 1 & 0 & 0 \\
\sin(\beta_0) & 0 & \cos(\beta_0) & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\]

(6.18)
\[
R_v(-\gamma_0) = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & \cos(\alpha_0) & \sin(\alpha_0) & 0 \\
0 & -\sin(\alpha_0) & \cos(\alpha_0) & 0 \\
0 & 0 & 0 & 1
\end{pmatrix}
\] (6.19)

Since the camera is attached to the local avatar and only rotation about the local \(y\) axis is allowed, the overall view matrix can be simplified by replacing \(\alpha_0\) and \(\gamma_0\) with zero if desired; however, to support rotation about all axes these variables are retained. The same condition is applied to the model matrix explained in Section 6.5 and a generic model can be achieved by extending the transformations and adding the rotation matrices to the overall transformation.

The simplified overall view matrix which results is as follows:

\[
\begin{pmatrix}
\cos(\beta_0) & 0 & -\sin(\beta_0) & z_0 \sin(\beta_0) - x_0 \cos(\beta_0) \\
0 & 1 & 0 & -y_0 \\
\sin(\beta_0) & 0 & \cos(\beta_0) & -z_0 \cos(\beta_0) - x_0 \sin(\beta_0) \\
0 & 0 & 0 & 1
\end{pmatrix}
\] (6.20)

### 6.7 Perspective projection

The perspective projection is exploited to transform the calculated positions of the blocks from eye coordinates into their pixel coordinates on the 2D screen. By looking at the graphics pipeline, the projection of a block onto the screen can be assessed (Figure 6.4).

As demonstrated in Figure 6.4, each block is passed through this pipeline and the block is multiplied by various matrices and then clipped if necessary. If the block successfully passes the clipping stage, it is ultimately mapped onto the screen.

As shown in the pipeline, the modelling and viewing transformation can be replaced by a single \textit{ModelView} matrix, which is in fact the product of model
Figure 6.4 Graphics pipeline

and view matrices (VM).

$$VM = \begin{bmatrix}
\cos(\beta) \cos(\beta_0) + \sin(\beta) \sin(\beta_0) & 0 & \cos(\beta_0) \sin(\beta) - \cos(\beta) \sin(\beta_0) & A \\
0 & 1 & 0 & B \\
\cos(\beta) \sin(\beta_0) - \sin(\beta) \cos(\beta_0) & 0 & \cos(\beta_0) \cos(\beta) + \sin(\beta) \sin(\beta_0) & C \\
0 & 0 & 0 & 1
\end{bmatrix}$$

(6.21)

$$A = z_0 \sin(\beta_0) - x_0 \cos(\beta_0) + \cos(\beta_0) \left( x - \frac{A_z \sin(\beta)}{2} + A_x \cos(\beta) \frac{w_m}{2} - \frac{w_i}{2} + x_m \right)$$

$$+ \sin(\beta_0) \left( \frac{A_z \cos(\beta)}{2} - z + A_x \sin(\beta) \frac{w_m}{2} - \frac{w_i}{2} + x_m \right)$$

$$B = y - y_0 - A_y \frac{h_m}{2} - \frac{h_i}{2} + y_m$$

$$C = \sin(\beta_0) \left( x - \frac{A_z \sin(\beta)}{2} + A_x \cos(\beta) \frac{w_m}{2} - \frac{w_i}{2} + x_m \right) - x_0 \sin(\beta_0) - z_0 \cos(\beta_0)$$

$$- \cos(\beta_0) \left( \frac{A_z \cos(\beta)}{2} - z + A_x \sin(\beta) \frac{w_m}{2} - \frac{w_i}{2} + x_m \right)$$
The projection matrix scales and translates each block in a way that all the blocks located inside the view frustum bounded by near and far clipping planes with the angle $fovy$ will lie in a standard cube that extends from $-1$ to $1$ in each dimension. After applying the perspective division step, the blocks in normalised device coordinates are multiplied by the view port matrix and transformed to the window coordinates. Then a simple mapping process maps the values to pixel coordinates.

\[
P = \begin{pmatrix}
\frac{1}{asp \times \tan(fovy)} & 0 & 0 & 0 \\
0 & \frac{1}{\tan(fovy)} & 0 & 0 \\
0 & 0 & \frac{far + near}{near - far} & \frac{2 \times far \times near}{near - far} \\
0 & 0 & -1 & 0
\end{pmatrix}
\] (6.22)

$P$ represents the perceptive projection matrix, in which $fovy$ is the vertical viewing angle of the camera and $asp$ is the aspect ratio, which controls the horizontal viewing angle relative to the vertical viewing angle. In other words $fovx = asp \times fovy$. The parameters near and far represent the inverted $z$ coordinates of the near and far clipping planes respectively.

### 6.8 Block behaviour

Using the transformation matrices obtained in the previous sections, the corner points of the blocks can be transformed to screen coordinates and hence width and height of each block can be calculated.

In the following sections, different scenarios are constructed to study the impact of virtual distance and orientation of video surface and camera on the projected blocks’ width and height. The behaviour of the projected blocks’ sizes is then used to confirm that different degrees of spatial degradation are required for different spatial regions of a texture.

Since various situations are analysed in the following experiments, textures are partitioned to blocks of $64 \times 64$ pixels, so that each spatial region represents a
distinct behaviour.

6.8.1 Characteristics of the rendered view of the clients terminal

In the experiments, the IVC window has a fixed resolution of 1152×864 pixels. The camera is located at the origin and its view frustum is aligned with the z axis. The near and far clipping planes are located at 0.3 and 1000 metres respectively and camera has a 60 degree vertical viewing angle (fovy) and 80 degree horizontal viewing angle (fovx) with aspect ratio of 4 : 3. The size of the video surface is configured such that its projection on the screen at 1 metre distance (0, 0, 1) from the camera eye and zero Euler angles (α₀ = 0, β₀ = 0, γ₀ = 0) has a resolution of exactly 512×512 pixels. This situation is called the generic situation of the IVC.

6.8.2 Block behaviours at different virtual distances

In this experiment, the camera is located at (0, 0, 0) and oriented in the direction (0, 0, 1). The position of video surface is varied from (0, 0, 1) to (0, 0, 10) in increments of 1 metre without changing its orientation. The height and width of all the blocks’ projections on-screen are calculated for each position.

Since rotation usually causes uninformed distortions and neither video surface nor the camera is rotated in this experiment, all the blocks have the same size in each 3D situation. However, as shown in Figure 6.5 and 6.6, due to the perspective projection, the further video surface is located, the smaller the height and width of the block will be. In other words, the size of blocks has an inverse relationship with the distance of video surface (z value).

6.8.3 Presentation of blocks

In all experiments, if the orientation of an avatar is changed, it is rotated about the local y axis. Therefore, the blocks in a single column will remain approximately the same size as each other. Thus, the maximum height and width of the blocks in a column are shown in Figure 6.7. Representative blocks from
Figure 6.5 Impact of virtual distance of video surface on the blocks’ projected width

Figure 6.6 Impact of virtual distance of video surface on the blocks’ projected height
the leftmost to rightmost columns are denoted Block1 to Block8 respectively; each BlockN has the same maximum width and height as all the blocks in that column.

### 6.8.4 Block behaviours with different virtual orientations

The video surface in this experiment is initially located at $(0, 0, 1)$ and initially oriented towards the camera (located at $(0, 0, 0)$ and oriented towards $(0, 0, 1)$). The orientation of the surface is changed gradually; initially the surface faces toward the camera, then it is rotated about the local $y$ axis in steps of $10^\circ$ until $\beta$ reaches a relative angle of $90^\circ$ with respect to the camera’s orientation.

The change in width of each of the eight blocks with respect to $\beta$ exhibits significantly different behaviour to the change in height with respect to $\beta$ due to the orientation of the axis of rotation. The height of the blocks varies more linearly, while the width of the blocks varies in a manner which is more strongly dependent on the distance from the $y$ axis.

Generally, the blocks’ behaviours can be categorised in three sections: the blocks
Figure 6.8 Impact of virtual orientation of video surface on the blocks’ projected width at a distance of one metre

located at the left side of the local $y$ axis, the blocks located near the $y$ axis and the blocks at the right side of the $y$ axis.

All the blocks’ widths are related to $\cos(\beta)$, thus they should have their maximum value at $\beta = 0^\circ$ and a minimum of zero $\beta = 90^\circ$. The blocks on the axis have the expected behaviour. However, as shown in Figure 6.8, width of the blocks located at the right side of the local axis are magnified first and then shrunk gradually and widths of the blocks at the left side of the axis are decreased with a higher rate. These behaviours are the impact of perspective projection and the fact that due to the rotation, the right side of video surface moves toward the camera and left side goes further away from the camera. In this experiment the surface is located fairly close to the camera, so this change of distance has a noticeable impact. However, if video surface is positioned further away, this impact is greatly reduced (Figure 6.10 and 6.11). The behaviour of the blocks’ heights are also affected by the distance change caused by the rotation. However, the rotation about $y$ axis has no impact on them directly, since points with the same $x$ value are parallel to the local $y$ axis. Hence, the
6.8.5 Block behaviours with different camera orientations

In this experiment, the video surface remains at (0, 0, 1) oriented towards (0, 0, 0) while the orientation of the camera located at (0, 0, 0) is varied. Since the camera’s horizontal viewing angle ($fovx$) is 80°, the orientation of the camera is varied from -40° to 40° in 10° increments. Next, to analyse the impact of distance at the same time, video surface is moved away from the camera gradually to reach a virtual distance of 5 metres from the camera.

Figure 6.12 and 6.13 demonstrate the impact of camera’s orientation on widths and heights of the projected blocks respectively. The results show that the camera’s orientation has a huge impact on the projected blocks’ sizes. The more the camera is rotated, the further video surface is pushed to the side of the

---

1Values outside of this range are impractical, since the video surface would not be inside the view frustum and it will be culled, hence the blocks would not be projected on the screen.
Figure 6.10 Impact of virtual orientation and distance of video surface (distance 2 to 5 metres)
Perceptual Pruning

Figure 6.11 Impact of virtual orientation and distance of video surface (distance 6 to 9 metres)
Figure 6.12 Impact of virtual orientation of the camera on the blocks’ projected width at 1 metre distance

Figure 6.13 Impact of virtual orientation of the camera on the blocks’ projected height at 1 metre distance
screen and the closer the orientation of the surface becomes to the orientation of the side plane of the view frustum. Consequently, blocks are enlarged as they move to the side of the screen - particularly when the avatar is in close proximity to the camera. However, in the extreme angles (-40 and 40), part of the surface might be outside of the view frustum and hence those blocks with the greatest increase in size are not projected on the screen (Figure 6.14).

Regardless of the virtual distance of video surface, the camera’s orientation affects the size of the blocks significantly. The outcomes for a distance of two to five metres are presented in Figure 6.15. However, the further away the surface is located, the sooner it exits the view frustum due to the rotation of the camera, limiting the impact at greater distances.

6.8.6 Characteristics of video surface

In practical situations, a video surface and camera may have any positions and orientations within the virtual environment. Since avatars are projected from the local client’s perspective, the quality of videos should be adjusted based on the sizes of the projected blocks. The whole system can therefore be characterised with respect to the camera’s position and orientation. Hence video surfaces with any position and orientation in the environment can be defined with three parameters uniquely (Figure 6.16). These parameters are
Figure 6.15 Impact of virtual orientation of camera and distance of video surface (distance 2 to 5 metres)
Figure 6.16 Characterisation of a video surface

the variables studied in Section 6.8.2 to 6.8.5.

The relative distance of video surface to the camera may be obtained from Equation 6.23 and is denoted by $d$. The relative angular distance of the camera from the normal vector from the video surface is $\beta$, the Euler angle about the local $y$ axis. $\beta_0$ represents the relative angular distance of the centre of the video surface with respect to the camera’s normal vector. The visual presentation of the scene and hence the projection of video surface are exactly the same if the camera is rotated by $30^\circ$ about its local $y$ axis or the video surface is rotated by $-30^\circ$ about the camera’s local $y$ axis. Hence, avatars with any positions in the scene can be presented by a relative angle of $\beta_0$ which can be calculated with Equation 6.24.

\[
d = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}
\]  
(6.23)
where \( x_0, y_0 \) and \( z_0 \) represent the Cartesian coordinates of the camera.

\[
\beta_0 = \beta_c - \left( \tan^{-1} \left( \frac{x}{z} \right) \right)
\]

(6.24)

where \( \beta_c \) is the absolute Euler angle about the \( y \) axis of the camera.

To have a more realistic scenario, the impact of simultaneous variation of key parameters (\( \beta \) and \( \beta_0 \)) on projection sizes of the blocks is demonstrated in Figure 6.17. Note that in this experiment distance (\( d \)) of video surface to the camera is constant and equal to 1.

In Section 6.8.4 and 6.8.5, the impacts of virtual orientation of video surface and camera are investigated separately. The joint impact of these parameters on the projected blocks’ sizes is analysed in this section. The outcome shows that the huge expansion of the projected blocks’ width and height caused by camera rotation reinforces the enlargement of blocks’ sizes achieved solely by rotating the video surface (Figure 6.10 and 6.11).

As expected the blocks’ projection sizes do not have the symmetric behaviour as if they would have, without introducing the camera rotation. For instance, the neutral 3D situation where the projected blocks’ sizes have neither expansion nor compression, which was obtained when the video surface was at its generic situation \(^2\), is achieved at 10, 20 and 30 degrees rotation of the video surface about its local \( y \) axis, when the camera is rotated -10, -20 and -30 degrees about the camera’s local \( y \) axis respectively. This shows the complexity of the combination of situations that may be encountered in a realistic situation.

Although, the only way of having a generic model that can handle full range of motion and any arbitrary sized spatial region is the matrix transformation of blocks as presented previously, the projection sizes of fixed sized spatial regions (e.g. blocks of \( 64 \times 64 \) or \( 8 \times 8 \) pixels), when the virtual motion is limited (i.e. avatars and the camera can only be rotate about their local \( y \) axes) can

\(^2\)Video surface is located at (0, 0, 1) and oriented towards (0, 0, 0), when the camera is located at the origin and oriented towards (0, 0, 1)
Figure 6.17 Impact of virtual orientation of camera and video surface on projection sizes of the blocks (at a distance of one metre)
be precomputed for combinations of virtual situations and used in real-time to minimise the computational cost of the projection calculation process. By this approach, the maximum bitrate saving might not be achieved but the computational cost is minimised. In other words, for thin clients that cannot handle the intensive process of transformation calculations, the performance of the mechanism in terms of reducing the required network capacity can be sacrificed for reducing the computational cost. In this approach, the virtual state of each avatar (video surface) with respect to the camera can be characterised by these three parameters. Then, the estimated required spatial quality (the matrix of the perceptual pruning mask (refer to Section 6.10)) of each region (all the blocks in a column) can be retrieved from a lookup table based on the parameters.

6.9 Degradation process

As discussed in section 6.8, size of the projected blocks may be smaller than their actual sizes due to the virtual distance and/or orientation of video surface. The hypothesis of this research is that the blocks with smaller projection sizes compared to their actual sizes require lower spatial resolution in that exact 3D situation. In other words, the spatial resolution of each block can be adjusted according to the size of the projected block on the screen. Nevertheless, this process is supposed to be performed on the server side or by the transmitting peers, and blocks are actually video frames’ macroblocks. Spatial degradation of a frame requires decoding of the whole frame; unfortunately, this decoding process is computationally expensive and introduces delay (Figure 6.18). Hence, a simple DCT downsampling method is proposed that only involves partial decoding of the frame. In the degradation process only entropy decoding is performed and the frame’s data in frequency domain is retrieved. Then size of each projected block based on the 3D situation of video surface and camera is

---

3Since, the maximum size of the projected blocks in each column is used for all the blocks, the projection sizes of some blocks might be overestimated and hence the maximum saving might not be obtained. However, the increase is negligible, as the projected blocks’ sizes are very close in a column.
calculated and all the coefficients that lie outside of the projected boundaries are replaced with zeros.

6.9.1 Justification of the method

To achieve the spatial degradation, a frequency masking method is proposed that accomplishes the goal with a low computation cost and delay. In the proposed method, the size of the projected block is calculated and the vertical and horizontal frequencies outside of the projected boundaries are zeroed out. It is hypothesised that when the block is projected to a smaller spatial size, only the frequencies in the projected range are visually perceptible. Hence, if a block of $8 \times 8$ pixels is projected to a $6 \times 2$ pixel region, only the DCT coefficients in this region are maintained and the rest are replaced by zeros.

In the first experiment, a $512 \times 512$ pixel reference image (Lena) is degraded spatially in the pixel domain and also degraded based on the above-mentioned frequency masking method and the results are compared. The degradation in pixel domain is achieved by replacing the average value of every $4$ ($2 \times 2$) adjacent pixels with one pixel, hence width and height of the resulting image are half of the original sizes ($256 \times 256$). For the frequency-masking method, the DCT coefficients of $8 \times 8$ blocks of the image are calculated, and only the $4 \times 4$ top-left sub-matrix (i.e., the DC component and low AC frequencies components) are retained and the rest are discarded. Since the spatial size of the image is half
the original size, the DCT coefficients are quantised accordingly (divided by 2) and the inverse DCT is obtained and the 256×256 image is constructed.

The results show that the resulting images have minor errors only and they are very close in terms of perceptual quality (Table 6.1).

However, in IVC systems, the projection of each block is calculated and frequency masking is applied based on the projected sizes and the DCT coefficients are not discarded but replaced with zeros. Hence the resulting frame size is the same as the original image size (512×512), while the bitrate is significantly reduced (especially for the I-frames).
6.10 Perceptual pruning

The process of calculating sizes of projected blocks and masking DCT components outside of the projected boundaries to reduce the required downstream bitrate of clients in IVC is referred to here as perceptual pruning. In order to analyse the impact of perceptual pruning method, a controlled IVC environment is configured, in which the general 3D specifications of the system for all experiments are fixed (refer to Section 6.8.1) and then the performance of the method in terms of perceptual quality and required network downstream capacity is studied.

6.10.1 The analysis of applying perceptual pruning on a static frame

To simplify the experiments, the studies are conducted using static images, which demonstrate the impact of the method on I-frames. Thus, Lena’s image is perceptually pruned for particular 3D situations. The pruned image is then applied as a texture on video surface and PSNR and similarity index of the degraded image versus the reference image in the exact same 3D situation are calculated. The objective quality assessment results are given to demonstrate the performance of the method in terms of perceptual quality. On the other hand, the percentage of zero DCT coefficients with respect to the total number of coefficients in each experiment is demonstrated as an indicator of potential improvement of required downstream bitrate.

Although the experiments are conducted using a static image, to achieve results with as much relevance as possible to actual video codecs, the image is divided to spatial regions that are commonly used in video codecs. Macroblock partitions typically used in video codecs include 16×16, 16×8, 8×16 and 8×8. The 8×8 partitions can also be further subdivided into partitions of size 8×4, 4×8 and 4×4. The partitioning is performed based on the video content. For instance 4×4 blocks are well suited for regions of the input image with sophisticated motion, while coarser partitions, e.g. 16×16, are better suited for smooth
Table 6.2
Objective assessment of reference image vs. degraded images outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>36.5714</td>
<td>32.5334</td>
<td>0.881</td>
</tr>
<tr>
<td>3</td>
<td>36.5714</td>
<td>32.5334</td>
<td>0.881</td>
</tr>
<tr>
<td>4</td>
<td>120.6549</td>
<td>27.3493</td>
<td>0.7225</td>
</tr>
<tr>
<td>5</td>
<td>120.6549</td>
<td>27.3493</td>
<td>0.7225</td>
</tr>
</tbody>
</table>

Table 6.3
Objective assessment of reference image vs. degraded images outside of the IVC system (8 × 8 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>29.4984</td>
<td>33.4668</td>
<td>0.8913</td>
</tr>
<tr>
<td>3</td>
<td>53.1827</td>
<td>30.9071</td>
<td>0.8307</td>
</tr>
<tr>
<td>4</td>
<td>98.1433</td>
<td>28.2462</td>
<td>0.7469</td>
</tr>
<tr>
<td>5</td>
<td>98.1433</td>
<td>28.2462</td>
<td>0.7469</td>
</tr>
</tbody>
</table>

areas. However, the partitions of 8×8 and 4×4 are widely used in video codecs and found to be well suited for the experiments. Hence the experiments are presented for both cases of dividing the image into 8×8 and 4×4 blocks.

6.10.1.1 Impact of virtual distance

In the first experiment, the reference image is pruned for different virtual distances from the camera. Since the projection of video surface at a virtual distance of 1 metre to the camera and the actual size of the reference image are 512×512 pixels, the perceptual pruning method does not degrade the image in this particular 3D situation. However, this experiment is repeated for 2, 3, 4 and 5 metres distance from the camera, which is achieved by changing the z value of video surface respectively. To demonstrate the impact of the proposed method, the reference image and degraded images are first compared out of IVC (Table 6.2 and 6.3) and then applied as textures and compared in the corresponding 3D situations.
Table 6.4
Objective assessment of reference image vs. degraded images inside the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2.6985</td>
<td>43.8535</td>
<td>0.991</td>
</tr>
<tr>
<td>3</td>
<td>1.0113</td>
<td>48.1159</td>
<td>0.9972</td>
</tr>
<tr>
<td>4</td>
<td>0.6132</td>
<td>50.2885</td>
<td>0.9985</td>
</tr>
<tr>
<td>5</td>
<td>0.9278</td>
<td>48.4902</td>
<td>0.9982</td>
</tr>
</tbody>
</table>

Table 6.5
Objective assessment of reference image vs. degraded images inside the IVC system (8 × 8 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3.4856</td>
<td>42.742</td>
<td>0.9889</td>
</tr>
<tr>
<td>3</td>
<td>7.2412</td>
<td>39.5667</td>
<td>0.9835</td>
</tr>
<tr>
<td>4</td>
<td>3.7252</td>
<td>42.4533</td>
<td>0.9919</td>
</tr>
<tr>
<td>5</td>
<td>7.0978</td>
<td>39.6535</td>
<td>0.9887</td>
</tr>
</tbody>
</table>

For each case, the reference image and degraded image are transmitted to IVC and applied on video surface separately, and then video surface with the corresponding texture is extracted from the rendered scene. The results are compared by calculating the PSNR and SSIM values (Table 6.4 and 6.5).

The outcomes confirm that the frame (image) can be degraded according to the size of the projected blocks, which is dependent on the position of video surface in this particular experiment. As demonstrated in Table 6.4 and 6.5, 4×4 partitioning of the frame results in better perceptual quality inside the IVC system. This behaviour is the result of DCT downsampling method that zeros out the coefficients out of projected boundaries. Since the blocks are smaller, the DCT coefficients in a block represent a smaller spatial region of the frame (image) and even in the maximum degradation case, in which all the components except the DC component are zeroed out, still the average value of all the pixels lying in the block is maintained. This mechanism causes a better perceptual quality for smaller blocks. However, there is a trade-off between quality and
<table>
<thead>
<tr>
<th>Dis</th>
<th>No. Coef</th>
<th>No. Zero Coef</th>
<th>Percentage before</th>
<th>Percentage after</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>262144</td>
<td>213</td>
<td>213</td>
<td>0.081253</td>
</tr>
<tr>
<td>2</td>
<td>262144</td>
<td>213</td>
<td>196610</td>
<td>0.081253</td>
</tr>
<tr>
<td>3</td>
<td>262144</td>
<td>213</td>
<td>225282</td>
<td>0.081253</td>
</tr>
<tr>
<td>4</td>
<td>262144</td>
<td>213</td>
<td>245760</td>
<td>0.081253</td>
</tr>
<tr>
<td>5</td>
<td>262144</td>
<td>213</td>
<td>245760</td>
<td>0.081253</td>
</tr>
</tbody>
</table>

**Table 6.6** Zero coefficients vs. non-zero coefficients (8 × 8 macroblocks)

<table>
<thead>
<tr>
<th>Dis</th>
<th>No. Coef</th>
<th>No. Zero Coef</th>
<th>Percentage before</th>
<th>Percentage after</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>262144</td>
<td>3229</td>
<td>3229</td>
<td>1.231766</td>
</tr>
<tr>
<td>2</td>
<td>262144</td>
<td>3229</td>
<td>196759</td>
<td>1.231766</td>
</tr>
<tr>
<td>3</td>
<td>262144</td>
<td>3229</td>
<td>196759</td>
<td>1.231766</td>
</tr>
<tr>
<td>4</td>
<td>262144</td>
<td>3229</td>
<td>245760</td>
<td>1.231766</td>
</tr>
<tr>
<td>5</td>
<td>262144</td>
<td>3229</td>
<td>245760</td>
<td>1.231766</td>
</tr>
</tbody>
</table>

**Table 6.7** Zero coefficients vs. non-zero coefficients (4 × 4 macroblocks)

Bitrate. Smaller partitions restrict the number of coefficients that can be zeroed out. For instance the maximum degradation, when the block size is 8 × 8 happens when only the DC component is kept and other 63 AC components are replaced with zeros, while this amount is decreased to 15 coefficients in the case of 4 × 4 block.

The number and percentage of zero coefficients for different virtual distances, which is an indication of the amount of required network capacity is demonstrated in Table 6.6 and 6.7.

6.10.1.2 Impact of virtual orientation

In this experiment, the camera is located at (0, 0, 0) and oriented towards (0, 0, 1), such that the camera is aligned with the z axis. The video surface is located at (0, 0, 1), while its orientation is varied. The perceptual pruning mechanism is performed for 0, 30, 60 and 80 degree rotation of video surface about the local y axis (β) and the resulting images are compared outside and inside of the IVC system. Like the previous experiment, two different spatial
<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>1.7773</td>
<td>45.6671</td>
<td>0.9927</td>
</tr>
<tr>
<td>60</td>
<td>13.158</td>
<td>36.9729</td>
<td>0.9582</td>
</tr>
<tr>
<td>80</td>
<td>47.6192</td>
<td>31.387</td>
<td>0.8779</td>
</tr>
</tbody>
</table>

Table 6.8
Objective assessment of reference image vs. degraded images outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>2.1359</td>
<td>44.869</td>
<td>0.9909</td>
</tr>
<tr>
<td>60</td>
<td>14.2281</td>
<td>36.6333</td>
<td>0.9516</td>
</tr>
<tr>
<td>80</td>
<td>49.2791</td>
<td>31.2382</td>
<td>0.8686</td>
</tr>
</tbody>
</table>

Table 6.9
Objective assessment of reference image vs. degraded images outside of the IVC system (8 × 8 macroblocks)

partitioning schemes (8×8 and 4×4) are examined.

The objective assessment results demonstrate the amount of degradation as shown in Table 6.8 and 6.9. The general result is that the more high frequency components are zeroed out, the worse the perceptual quality and the greater the error in the 2D presentation of the image outside IVC.

The experiment is repeated again, but this time the reference and degraded images are applied on video surface (figure 6.21). The results achieved from the 3D presentation of degraded images compared to the reference frames with the same angular states suggest a totally different behaviour and confirm our hypothesis. Despite the degree of DCT downsampling, the reference and degraded images have insignificant error and they are perceptually indistinguishable (Table 6.10 and 6.11).

Table 6.12 and 6.13 show the proportion of coefficients that are replaced with zero for each orientation. As expected, the larger the angle $\beta$, the more coefficients are zeroed out and more downstream capacity is saved.
Figure 6.20 Perceptual pruning for various virtual orientations (2D presentation)

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>0.8306</td>
<td>48.9707</td>
<td>0.9976</td>
</tr>
<tr>
<td>60</td>
<td>1.3789</td>
<td>46.7694</td>
<td>0.995</td>
</tr>
<tr>
<td>80</td>
<td>0.7742</td>
<td>49.2761</td>
<td>0.9982</td>
</tr>
</tbody>
</table>

Table 6.10
Objective assessment of reference image vs. degraded images inside of the IVC system (4 × 4 macroblocks)
**Perceptual Pruning**

Figure 6.21 Perceptual pruning for virtual orientations of 0 and 30 degrees (3D presentation)
Figure 6.22 Perceptual pruning for virtual orientations of 60 and 80 degrees (3D presentation)
Perceptual Pruning

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>1.9885</td>
<td>45.1796</td>
<td>0.9944</td>
</tr>
<tr>
<td>60</td>
<td>2.3307</td>
<td>44.4899</td>
<td>0.9918</td>
</tr>
<tr>
<td>80</td>
<td>1.8712</td>
<td>44.4899</td>
<td>0.9965</td>
</tr>
</tbody>
</table>

**Table 6.11**  
Objective assessment of reference image vs. degraded images inside of the IVC system (8 × 8 macroblocks)

<table>
<thead>
<tr>
<th>Angle</th>
<th>No. Coef</th>
<th>No. Zero Coeff</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>before</td>
<td>after</td>
<td>before</td>
</tr>
<tr>
<td>0</td>
<td>262144</td>
<td>213</td>
<td>213</td>
</tr>
<tr>
<td>30</td>
<td>262144</td>
<td>213</td>
<td>28052</td>
</tr>
<tr>
<td>60</td>
<td>262144</td>
<td>213</td>
<td>99680</td>
</tr>
<tr>
<td>80</td>
<td>262144</td>
<td>213</td>
<td>173362</td>
</tr>
</tbody>
</table>

**Table 6.12** Zero coefficients vs. non-zero coefficients (8 × 8 macroblocks)

<table>
<thead>
<tr>
<th>Angle</th>
<th>No. Coef</th>
<th>No. Zero Coeff</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>before</td>
<td>after</td>
<td>before</td>
</tr>
<tr>
<td>0</td>
<td>262144</td>
<td>3229</td>
<td>3229</td>
</tr>
<tr>
<td>30</td>
<td>262144</td>
<td>3229</td>
<td>20976</td>
</tr>
<tr>
<td>60</td>
<td>262144</td>
<td>3229</td>
<td>82714</td>
</tr>
<tr>
<td>80</td>
<td>262144</td>
<td>3229</td>
<td>159244</td>
</tr>
</tbody>
</table>

**Table 6.13** Zero coefficients vs. non-zero coefficients (4 × 4 macroblocks)
6.10.2 The analysis of applying perceptual pruning on a video stream

In this section, the perceptual pruning mechanism is applied on an actual video stream and the video quality and bitrate in different 3D situations are analysed. To make the outcome consistent with the previous results, the controlled IVC environment is configured as explained in Section 6.8.1. However, a modification is applied to make the IVC compatible with the used video stream. Since, the native spatial resolution of the utilised video stream is $352 \times 288$ pixels (CIF), the IVC setting is modified such that the projection of the video surface on the screen at 1 metre distance from the camera with zero Euler angles has the exact same CIF resolution of $352 \times 288$ pixels.

6.10.2.1 Video codec and video stream specifications

A H.264 video codec is adopted for this study. The perceptual pruning mechanism is then integrated with the code. The proposed downsampling method is exploited to reduce the raw data rate in addition to the basic encoding process. After calculating the motion vectors, the frames are partitioned into $4 \times 4 \times 4$ pixels blocks and an integer transform is applied on the data. The output of the transform is first quantised, then passed to the perceptual pruning component. The component computes a frequency mask based on the 3D situation of the video surface and the spatial position of the block. Then, the mask is applied on the quantised data. Finally, entropy encoding is applied to the masked values. The combination of the quantisation and perceptual pruning may zero out some coefficients, and the entropy coding combines a number of consecutive zero-valued coefficients with the value of the next non-zero quantised coefficient into a single symbol. Other mechanisms such as run length coding are also utilised to reduce the final bitrate.

In this experiment, the quantisation parameter of the codec is set to 21 and the motion search range is restricted to $\pm 8$ and $\pm 4$ pixels for the first and second

\footnote{Note that this partitioning is independent of the motion estimation calculation and any arbitrary sized block could be picked in the search process.}
Table 6.14
The average objective assessment results of reference frames vs. quantised frames outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.871</td>
<td>40.52</td>
<td>0.99284</td>
</tr>
</tbody>
</table>

levels of full search motion estimation respectively. The performance of the perceptual pruning method is evaluated by using the popular sequence known as Foreman. The sequence is in CIF resolution (352 × 288) and only the luma or Y component is considered. The first hundred frames of the sequence, which represent a talking head video are coded at 30 Hz. The group of pictures (GOP) structure is set to IPPPPPPP, which indicates that the first frame is coded as an I-frame and each I-frame is followed by 7 P-frames.

When only the standard video coding procedure (without perceptual pruning mechanism) is performed on the sequence, the reconstructed frames are called quantised frames. If the perceptual pruning is also added to the process, the achieved frames are called pruned frames. Note that the bitrate of the data is always analysed, when the perceptual pruning is involved.

The bitrate of the coded data is measured as well as the quality of the reconstructed frames. The objective quality is achieved by calculating the MSE, PSNR and SSIM of the quantised, pruned and reference frames versus each other. It should be considered that the metrics at this stage only represent the objective quality of the frame outside the IVC and do not reflect the performance of perceptual pruning mechanism. The quality of the frames are also assessed inside IVC after applying on the video surface. Objective metrics like PSNR and SSIM are used to evaluate the perceptual quality of the pruned frame versus the decoded and reference frames at different 3D situations.

6.10.2.2 Impact of virtual distance

In this experiment, the first 100 frames of the Foreman sequence are coded for 10 different virtual distances. Since, the IVC is configure in a way such that the
Table 6.15
The average objective assessment results of reference frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.871</td>
<td>40.52</td>
<td>0.99284</td>
</tr>
<tr>
<td>2</td>
<td>59.792</td>
<td>30.432</td>
<td>0.94725</td>
</tr>
<tr>
<td>3</td>
<td>59.792</td>
<td>30.432</td>
<td>0.94725</td>
</tr>
<tr>
<td>4</td>
<td>227.35</td>
<td>25.399</td>
<td>0.86491</td>
</tr>
<tr>
<td>5</td>
<td>227.35</td>
<td>25.399</td>
<td>0.86491</td>
</tr>
</tbody>
</table>

Table 6.16
The average objective assessment results of quantised frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Inf</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>41.352</td>
<td>32.063</td>
<td>0.94252</td>
</tr>
<tr>
<td>3</td>
<td>41.352</td>
<td>32.063</td>
<td>0.94252</td>
</tr>
<tr>
<td>4</td>
<td>245.28</td>
<td>24.896</td>
<td>0.85798</td>
</tr>
<tr>
<td>5</td>
<td>245.28</td>
<td>24.896</td>
<td>0.85798</td>
</tr>
</tbody>
</table>

projection of the video surface on the screen at 1 metre distance from the camera has the exact same CIF resolution of 352 × 288 pixels as the reference frame, the whole sequence is first coded for 1 metre distance and zero Euler angles (0, 0, 0). Then, the process is repeated for distances of 2 to 10 metres in increments of one metre without changing the orientation. The average objective results of the reconstructed frames outside of the IVC system is presented in Table 6.15 and Table 6.16.

As expected, the PNSR and SSIM values decrease as the virtual distance increases. This behaviour is due to the measurement of the objective metrics outside of IVC and the calculated mask based on the virtual distance. As explained in Section 6.8.2, due to the perspective projection the blocks’ sizes have an inverse relationship with the virtual distance. Hence, the further away the video surface is located, the smaller the projection of the blocks are and therefore, the more coefficients are located outside of the projected blocks’ boundaries.

5 Distances more than 5 metres are not presented, since they have the same values as distance 4.
and consequently replaced with zeros\textsuperscript{6}.

As a result, the bitrate of the coded video stream is reduced as well. However, it is observed that the objective assessment results and bitrate values are steady for virtual distances more than 4 metres. The maximum bitrate saving based on the proposed downsampling mechanism is achieved, when all the AC components are zeroed out and only the DC component is maintained. Due to the size of the blocks (4 × 4 pixels), all the 15 AC components are replaced with zero at distance of 4 metres and hence no more bitrate reduction can be obtained even at higher distances. Therefore, the MSE, PSNR, SSIM and bitrate values of the sequence at virtual distances more than 4 metres have the same value as the virtual distance of 4 metres.

The total and average bitrate of the coded sequence at different virtual distances are demonstrated in Figure 6.23 and 6.24 respectively. An average bitrate represents the average size of the coded frames, when the video surface has a specific virtual situation (i.e. at a certain virtual distance), while the total bitrate value shows the number of allocated bits in the output stream of the coded frames in one second duration of the sequence, when the video surface has a specific virtual situation.

Due to the existence of two types of frames in the coded sequence, the total and average bitrate values of the I and P frames are measured and shown separately in Figure 6.25 to 6.28.

Since, in this experiment the video stream is coded at 30 Hz, 4 I frames and 26 P frames exist in one second duration of the sequence. Thus, the total bitrate of the system in one second at any virtual situation has roughly the same proportion to the average I and P frame bitrate values at the same virtual situation.

The bitrate saving achieved by utilising the perceptual pruning mechanism is

\textsuperscript{6}Note that the presented PSNR and SSIM values are measured outside of IVC and the actual performance of the proposed method should be evaluated, when the objective assessment is conducted inside the IVC system.
Figure 6.23 Total bitrate of the video stream at different virtual distances

Figure 6.24 Average bitrate of a video frame at different virtual distances
Figure 6.25 Total bitrate of I-frames during one second at different virtual distances

Figure 6.26 Total bitrate of P-frames at different virtual distances
Figure 6.27 Average bitrate of an I-frame at different virtual distances

Figure 6.28 Average bitrate of a P-frame at different virtual distances
Figure 6.29 Total bitrate saving achieved at different virtual distances
demonstrated in Figure 6.29 to 6.31. The results show that the perceptual pruning reduces the total bitrate of the video data by 64.25% at virtual distance of 4 metre or more. The saving for I and P frames are in average 56.63% and 65.7% at distant situations respectively.

To evaluate the impact of perceptual pruning at different virtual distances, the reference, quantised and perceptually pruned sequences are applied on the video surface at the mentioned virtual distances and the objective quality of the degraded sequences are measured. The MSE, PSNR and SSIM values are presented in Table 6.17 and 6.18.

The difference of values in Table 6.17 and 6.18 shows the huge of impact of video coding process like quantisation and motion compensation on the reconstructed frames and hence perceptual pruning. As the PSNR is very sensitive to the created artifacts by the codec, the SSIM reflects the perceptual pruning performance better. Nevertheless, the results show that regardless of the virtual distance, the coded and pruned sequences have a very similar perceptual
Figure 6.30 Total bitrate saving per I-frames at different virtual distances

Figure 6.31 Total bitrate saving per P-frames at different virtual distances
Table 6.17
The average objective assessment results of reference frames vs. pruned frames inside of the IVC system (4 x 4 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.2627</td>
<td>39.015</td>
<td>0.98163</td>
</tr>
<tr>
<td>2</td>
<td>36.206</td>
<td>32.593</td>
<td>0.97658</td>
</tr>
<tr>
<td>3</td>
<td>35.046</td>
<td>32.735</td>
<td>0.98231</td>
</tr>
<tr>
<td>4</td>
<td>172.25</td>
<td>26.975</td>
<td>0.96984</td>
</tr>
<tr>
<td>5</td>
<td>170.54</td>
<td>27.022</td>
<td>0.97248</td>
</tr>
</tbody>
</table>

Table 6.18
The average objective assessment results of quantised frames vs. pruned frames inside of the IVC system (4 x 4 macroblocks)

<table>
<thead>
<tr>
<th>Distance</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Inf</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>21.158</td>
<td>34.948</td>
<td>0.97744</td>
</tr>
<tr>
<td>3</td>
<td>20.202</td>
<td>35.15</td>
<td>0.98347</td>
</tr>
<tr>
<td>4</td>
<td>182.67</td>
<td>26.369</td>
<td>0.96556</td>
</tr>
<tr>
<td>5</td>
<td>180.51</td>
<td>26.43</td>
<td>0.96861</td>
</tr>
</tbody>
</table>

quality inside the IVC system. Moreover, the pruned sequence is substantially smaller in terms of bitrate. Hence, it is confirmed that the perceptual pruning mechanism reduces the required network capacity with negligible perceptual impact on the clients.

6.10.2.3 Impact of virtual orientation

In this experiment, the camera is located at (0, 0, 0) and aligned with the z axis. The video surface is located at (0, 0, 1) and initially oriented towards the camera. Then the orientation of the video surface is changed to 0°, 30°, 60° and 80°. Consequently, The video sequence is coded and pruned for the mentioned orientations of the video surface (β). The perceptual quality of the sequence is assessed by the same objective metrics as Section 6.10.2.2. The assessment is performed outside and inside of the IVC system. Furthermore, The bitrate of the video sequence after applying the perceptual pruning is measured and the saving achieved by this mechanism is presented.
Figure 6.32 Process of calculating the perceptual quality of the video frame inside and outside the IVC system
Table 6.19
The average objective assessment results of reference frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.871</td>
<td>40.52</td>
<td>0.99284</td>
</tr>
<tr>
<td>30</td>
<td>30.655</td>
<td>33.31</td>
<td>0.99048</td>
</tr>
<tr>
<td>60</td>
<td>22.56</td>
<td>34.725</td>
<td>0.98439</td>
</tr>
<tr>
<td>80</td>
<td>70.265</td>
<td>29.79</td>
<td>0.94269</td>
</tr>
</tbody>
</table>

Table 6.20
The average objective assessment results of quantised frames vs. pruned frames outside of the IVC system (4 × 4 macroblocks)

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>1.0000</td>
</tr>
<tr>
<td>30</td>
<td>14.622</td>
<td>36.52</td>
<td>0.98965</td>
</tr>
<tr>
<td>60</td>
<td>12.045</td>
<td>37.438</td>
<td>0.98054</td>
</tr>
<tr>
<td>80</td>
<td>56.827</td>
<td>30.926</td>
<td>0.93813</td>
</tr>
</tbody>
</table>

The measured MSE, PSNR and SSIM values of the reference and quantised sequences versus the pruned sequence for different virtual orientation are given in Table 6.19 and 6.20. Note that the values are measured outside of the IVC system.

The objective assessment is performed, when the sequences are imported to the IVC system and applied on the video surface. The outcomes are presented in Table 6.21 and 6.22.

The bitrate of the pruned sequence at 0° to 90° virtual orientation of the video

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8.2627</td>
<td>39.015</td>
<td>0.98163</td>
</tr>
<tr>
<td>30</td>
<td>26.961</td>
<td>33.86</td>
<td>0.9778</td>
</tr>
<tr>
<td>60</td>
<td>17.336</td>
<td>35.837</td>
<td>0.98282</td>
</tr>
<tr>
<td>80</td>
<td>26.615</td>
<td>33.915</td>
<td>0.99086</td>
</tr>
</tbody>
</table>

Table 6.21
The average objective assessment results of reference frames vs. pruned frames inside of the IVC system (4 × 4 macroblocks)
Perceptual Pruning

<table>
<thead>
<tr>
<th>Angle</th>
<th>MSE</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Inf</td>
<td>1.0000</td>
</tr>
<tr>
<td>30</td>
<td>15.834</td>
<td>36.188</td>
<td>0.97905</td>
</tr>
<tr>
<td>60</td>
<td>9.4818</td>
<td>38.416</td>
<td>0.98328</td>
</tr>
<tr>
<td>80</td>
<td>19.607</td>
<td>35.436</td>
<td>0.99181</td>
</tr>
</tbody>
</table>

Table 6.22
The average objective assessment results of quantised frames vs. pruned frames inside of the IVC system (4 × 4 macroblocks)

The surface are measured and demonstrated in Figure 6.33 and 6.34.

The average and total bitrate of the sequence after applying the perceptual pruning mechanism at different virtual orientations are separately presented for I and P frames in Figure 6.35 to 6.38.

The outcome confirms that the bitrate of the pruned sequence reduces as the virtual orientation increases. This reduction is obtained due to the compression of the projected blocks’ sizes, when the video surface is rotated.

The network capacity saving achieved at different virtual orientation, when the video surface is located at 1 metre distance is presented in . The saving per I and P frames are shown in Figure 6.39 to 6.41.

The results revealed that the downstream bitrate is respectively decreased by 39.22%, 34.52% and 40.11% for the whole sequence, I frames and P frames, when the orientation of the video surface is at 90°. Please note that sub pixel degradation is not considered in the perceptual pruning mechanism. Thus, in order to avoid any perceptual impact on the quality of the pruned frame, the projection of the blocks are over calculated and the DC component is maintained in any situation. Hence, in the situation that the projected block is distorted the maximum bitrate saving might not be achieved. Therefore, a higher amount of savings are observed in the situations that only distance of the video surface is increased and the orientation of the video surface is maintained toward the camera.
Figure 6.33 Total bitrate of the video stream at different virtual orientations

Figure 6.34 Average bitrate of a video frame at different virtual orientations
Figure 6.35 Total bitrate of I-frames at different virtual orientations

Figure 6.36 Average bitrate of an I-frame at different virtual orientations
Perceptual Pruning

Figure 6.37 Total bitrate of P-frames at different virtual orientations

Figure 6.38 Average bitrate of a P-frame at different virtual orientations
**Figure 6.39** Total bitrate saving achieved at different virtual orientations

**Figure 6.40** Total bitrate saving per I-frames at different virtual orientations
6.10.2.4 Joint impact of virtual distance and orientation

In this experiment, the virtual distance and orientation of the video surface are changed simultaneously and the quality and bitrate of the sequence are analysed. In the whole experiment, the camera is located at the origin and aligned with z axis. The video surface is moved further to (0, 0, 2) and the orientation of the surface is varied from 0° to 90° in increments of 10°. Then, the perceptual pruning is performed for the exact 3D situation of the video surface and the resulting sequence is compared outside and inside of the IVC system with the reference and normally coded (without perceptual pruning) sequences, respectively.

The objective assessment results demonstrate the amount of degradation applied to the image by the perceptual pruning. The more high frequency components are zeroed out, the worse the perceptual quality becomes in the 2D presentation of the sequence outside of IVC. However, when the sequences are
applied on the video surface inside IVC, the error is minimal and the sequences have negligible perceptual difference in their qualities.

The bitrate of the sequence pruned for different virtual orientations at distance of 2 metres shows a different behaviour from the pruned sequence at virtual distance of 1 metre. As demonstrated in Figure 6.42, the bitrate is increased at 10° rotation and this increase is reduced moderately at 20° and maintained until \( \beta \) reaches 50°. Nevertheless, the bitrate is still higher than 0 degree orientation in the entire range (0° to 50°). Then, the bitrate is decreased when the video surface is rotated more than 50° as expected. This behaviour is due to the expansion of the blocks (texture magnification) explained in Section . It is expected that the change of blocks’ widths has a relationship with \( \cos(\beta) \). So they have their maximum and minimum values at \( \beta = 0° \) and \( \beta = 90° \), respectively. However, the blocks located at the right side of the axis (the side that moves toward the camera) are magnified, when the video surface is rotated (has angles more than 0°). This increase in the bitrate is due to this expansion which is result of the perspective project. The reason that this behaviour is not observed at 1 metre distance, is that the video quality is at its highest resolution and higher resolutions are not available, hence the magnification is ignored.

Note that the height of the blocks are also increased in the same range. Nevertheless, they are influenced by the distance change due to the rotation (the right side moves toward the camera, when the surface is oriented at positive \( \beta \) angles) and not directly caused by the rotation about the y axis as the points with the same x values are parallel to the local y axis and hence not affected.

As presented in Figure 6.44, when then surface is located further away (\( \geq 3 \) metres), this impact is significantly reduced. Since, the impact of distance change caused by the rotation becomes less effective as the surface moves away. Consequently, only the direct impact of rotation, which has a direct relationship with \( \cos(\beta) \) is observed. Although the bitrate is increased in virtual distance 2, the absolute bitrate is still much lower (as seen in Figure 6.44). Hence, perceptual pruning never causes an absolute increase in the bitrate.
Figure 6.42 Total bitrate of the video stream at different virtual orientations

Figure 6.43 Average bitrate of a video frame at different virtual orientations
Figure 6.44 Total bitrate of the video sequence at different virtual distances and orientations

The total and average bitrate of the sequence at different virtual orientations, when the video surface is located at 3 metres to 4 metres from the camera are presented in Figure 6.45.

Since the sequence is consisted of I and P frames, the total and average bitrate of each frame type, when the virtual distance varies from 2 to 4 and virtual orientation is increased from 0° to 90° in steps of 10° are demonstrated at Figure 6.46 and 6.47.

The bitrate saving achieved in the studied 3D situations are given in Figure 6.48 and 6.49.

6.11 Simulation results

In the simulations, the IVC is configured as described in Section 6.8.1. The local avatar with a random orientation is located on the origin of a 20×20
Figure 6.45 Total and average bitrate of the sequence at different virtual orientations (virtual distance of 3 and 4 metres)
Figure 6.46 Total bitrate of I and P frames at different virtual orientations (virtual distance of 2 to 4 metres)
Figure 6.47 Average bitrate of I and P frames at different virtual orientations (virtual distance of 2 to 4 metres)
Figure 6.48 Total bitrate saving of the video sequence at different virtual orientations (virtual distance of 2 to 4 metres)
Figure 6.49 Total bitrate saving per I and P frames at different virtual orientations (virtual distance of 2 to 4 metres)
metres flat environment. Furthermore, all simulations are performed based on 100 iterations.

Since the performance of perceptual pruning mechanism was analysed based on Lena’s image and a video stream and the results showed that the worst performance of the mechanism is obtained, when I frames are pruned, the total size of I-frames in the required video streams, when Lena’s image is the I-frame, is measured to present the worst case scenario. This measurement is performed for a chunk of data with only one I-frame per client in the given time. In the actual implementation of the perceptual pruning mechanism, when an I-frame is perceptually pruned, the same degradation is applied to all subsequent P-frames as well. Note that all the simulation settings are as the previous chapters.

6.11.1 Impact of density

In this section, the impact of density on the total size of I-Frames is studied. The number of avatars (video surfaces) is increased linearly from 5 to 60 and the avatars with random orientations are uniformly randomly distributed in the environment. Note that first the area of interest management mechanism proposed in Chapter 3 is utilised, and then the perceptual pruning method is applied on the avatars that survived the culling processes.

The outcome shows that the perceptual pruning method can improve the network downstream capacity saving in a crowded situation by an additional 56.58% on average. The total downstream bandwidth required for I-frames can be reduced by up to 96.27% by exploiting both the AOI and perceptual pruning mechanisms in an IVC system. However, the majority of this saving is achieved through the utilisation of AOI method.

6.11.2 Analysing a pathological scenario

In this simulation, the AOI management enhancements are avoided by establishing a pathological scenario. The worst case scenario for the AOI management would be one in which all the avatars are not only located in the viewing frus-
Perceptual Pruning

Figure 6.50 Impact of density on the total size of I-frames

tum of the viewer but also facing toward the camera and none of the avatars are blocked by others. This scenario can be simulated in a lecture theatre situation, where all the avatars are arranged in the viewing frustum of the local client in a way that no occlusion is happening. In this situation, the AOI mechanism is ineffective.

This simulation is performed for the same number of avatars as the previous simulation. The number of avatars in the lecture theatre is increased from 5 to 60 in increments of 5. The avatars are oriented to directly face the local avatar, which means no minification or magnification is caused by rotation about the local axes of video surfaces ($\beta$ impact).

The results (Figure 6.51)\textsuperscript{7} revealed that the perceptual pruning method in the lecture theatre situation can reduce the required downstream bandwidth for I-frames by the average of 61.24%.

\textsuperscript{7}Due to the pathological scenario the AOI mechanism is not effective in this situation and hence all the outputs of different culling techniques used in the AOI method have the same values and hence, the lines are overlapped
6.11.3 Analysing a realistic scenario

In this simulation, we attempted to avoid the use of the AOI method while making the scenario more realistic. Hence, the lecture theatre arrangement with the same number of avatars (5 to 60) is repeated. However, the avatars are not oriented strictly to face the local avatar, but they have a uniformly distributed random orientations in the range $(-90^\circ, 90^\circ)$ with respect to the camera. Thus, none of the clients would be culled due to AOI pruning, but they may be oriented at a relative angle to the camera. In the simulated scenario the impact of rotation about the local $y$ axis of the avatars is also considered (Figure 6.53).

Please note that in many situations inside the IVC system, when the participants form conversations or join one, it is observed frequently that avatars have steep angular orientations ($> 60^\circ$) and still in the centre of attention of others (e.g. Avatar B and C in Figure 6.52).

As discussed in Section 6.8.4, change of $\beta$ may cause minification as well as mag-
Figure 6.52 Avatars with steep angular situations in IVC

nification. Obviously, the minifications are utilised to reduce the downstream bitrate. Magnifications may result in smaller downstream bitrate savings, but ignoring them leads to a noticeable degradation. Note that magnifications never cause an absolute increase in the bitrate, since no upsampling is performed in this method. In the case that the avatar is located at a distance of one metre, which requires the highest resolution (512×512 in this study) with zero Euler angle $\beta$ about the local $y$ axis, magnifications are ignored, since higher resolutions are not available and no method can rectify this situation. In the other situations, the base resolution is already lower than the highest resolution. If the blocks are magnified in these situations, higher resolutions that are still lower or equal to the highest resolution are available and no absolute increase will occur, while the degradation remains imperceptible.

In this simulation, the perceptual pruning mechanism saves even more downstream bitrate (an average saving of 64.07%). This is due to the higher number of minifications compared to magnifications.

6.12 Conclusion

We proposed a novel video degradation mechanism known as perceptual pruning in this chapter. This method not only adjusts video quality of avatars based
on their virtual distances and orientations, but also applies different levels of degradations on different spatial regions of videos. The differentiation of video quality in spatial regions of a video delivers an optimum sized video stream with a high perceptual quality. Moreover, a mathematical model based on 3D transformations of objects in 3D environments is introduced. Using the mathematical model, the exact projection of any arbitrary sized and shaped spatial region of a texture can be calculated. The model is then adapted to utilise video macroblocks as the spatial regions. By exploiting the model, the size of the projected macroblocks on the screen is computed and the video stream is accordingly adjusted. To minimise the computational cost of the process, a DCT downsampling method is proposed that avoids complete decoding of video data. The downsampling method is applied after entropy decoding and benefits from a simple frequency masking mechanism that degrades the spatial quality of video macroblocks in the frequency domain. During the masking process, the frequencies lying outside of the projected boundaries of the block are zeroed out.
By using objective methods for assessing perceptual quality such as PSNR and SSIM, the performance of perceptual pruning in different 3D situations is evaluated. Having a similarly index of more than 0.99 and PSNR of more than 40 in majority of situations reveals the efficiency of the method in terms of perceptual quality. On the other hand, a combination of perceptual pruning and AOI methods can save up to 96.27% of the downstream bitrate of IVC in a crowded environment. Finally the method is challenged in a pathological scenario when the AOI mechanism is inapplicable. The outcomes demonstrated that perceptual pruning can reduce the total size of I-Frames by 64.07% on average, when the AOI mechanism is not functional.
Chapter 7

Conclusion

An Immersive video conferencing system (IVC) as described in this research employs 3D virtual environments to extend the visual space of video conferencing. It is shown that the 3D characteristics of IVC not only create a sense of immersion, but also help to overcome scalability barriers that the existing video conferencing systems are struggling with.

Due to the 3D presentation of IVC, limited number of avatars can be accommodated in the client’s view field at the time. However, this restriction is acceptable to the clients, since it mimics the known character of crowd interactions in the real world. By exploiting this attribute of IVC, an area of interest (AOI) management technique is proposed, which creates a conservative list of video streams of avatars that can potentially be visible. This list is updated constantly based on the client’s perspective. Consequently, only the necessary video streams are transmitted to the client and hence the required downstream bitrate is reduced. AOI management benefits from four methods. It is demonstrated in this thesis that by utilising the distance-based culling method in a crowded environment with uniform distribution, a modest bit rate saving of approximately 22% can be achieved. A considerable downstream bitrate saving of 73.66% is obtained after adding the view frustum culling technique. The back face culling method enhances the result to 85.58% by avoiding the transmission of the avatars that are facing away. Finally, by exploiting the occlusion culling mechanism in addition to other methods, total downstream bitrate of IVC is 175...
Conclusion

reduced by up to 90.61% compared to a typical video conferencing application.

Furthermore, the thesis investigates the pathological scenarios where the AOI mechanism is ineffective. A lecture theatre arrangement, for example, is presented in which not only all the avatars are in the visual range and view frustum of the viewer, but also they are facing toward him/her and none of them are occluded. It is demonstrated that in these situations, it is still possible to reduce the downstream bit rate by differentiating the video quality of avatars and transmitting lower video resolutions for the avatars located further away from the view point. Hence, a video quality differentiation (VQD) model is proposed that adjusts video quality of avatars with respect to their 3D situations relative to the viewer. To obtain this model, a video quality assessment in the context of a representative 3D IVC is conducted. A total of 233 individuals participated in the study and scored the video quality of avatars with various 3D situations. The model achieved by the results of the survey predicts the required video quality with an average Pearson correlation of 0.9998. It is shown that by employing the VQD mechanism in a lecture theatre situation, the average downstream bitrate can be decreased by 74.60% with negligible perceptual impact.

Nevertheless, the user study results revealed that the uniform degradation of video quality as exists in the state-of-the-art video quality differentiation models is inefficient, when the projection of the video surface is distorted (e.g. rotation of the surface or camera). Hence, a perceptual pruning method is developed and presented in this research, that takes the 3D transformations of video surface into account. Furthermore, the perceptual pruning method adapts the video coding’s partitioning and calculates the projection size of each block. Based on the projection sizes, the spatial resolution of each region is adjusted independently. This research also showed that due to the cheap computational cost of perceptual pruning, which only requires entropy decoding of the frames, the process can be carried out on the server side or peers before sending the video stream to the clients. The outcomes confirm that the combination of perceptual pruning and AOI mechanism can decrease the total downstream bitrate by
96.27% in a realistic scenario.

7.1 Future works

There are number of interesting areas for future research based on the work undertaken in this research. These include:

- **Subjective assessment of the perceptual pruning mechanism**: Although, the method is assessed by objective metrics like PSNR and SSIM that are widely accepted and have a good correlation with human perceived visual quality respectively. We intend to conduct a subjective assessment to analyse the sensitivity of human subjects to the proposed spatially differentiated degradation method.

- **Utilisation of predefined frequency masks in perceptual pruning mechanism**: There is always a trade off between accuracy and computation. Due to the required computation for calculation of the blocks’ projections, computation can become a bottleneck. In order to overcome this issue, the spatial regions of the video surface can set to be static (e.g. 64×64 blocks) and the virtual states can be limited to only critical situations (e.g. three spatial distances and six spatial orientations). Then, a number of frequency masks can be pre-calculated for the combination of all virtual distances and orientations. These masks can be employed on the server side based on a low bit rate feedback signal from the clients. Hence no calculation is required to be performed on the server side with respect to avatar transformations and the appropriate predefined masks are simply applied.

- **Integration of perceptual pruning with wavelet coding**: Wavelet coding provides scalability at the bitstream level, which can be utilised by perceptual pruning. DCT based codecs are not usually granular enough to provide many different resolutions at an affordable computational cost in real-time. The available solutions split the video stream into a num-
ber of sub-streams, known as layers or descriptions, which is suitable for heterogeneous population of recipients with limited number of video quality requirements. while, use of wavelet transformations provide a spatial scalability inherently. This feature can be incorporated with 3D transformation proposed in perceptual pruning to obtain a real-time granular degradation mechanism for IVC.
Bibliography


Appendix A

Experimental Settings and Subjects’ Demographics

A.1 Introduction

Measurement of perceptual video quality is of fundamental importance for any video conferencing system including IVC. Hence, a web-based subjective video quality assessment (VQA) was developed to evaluate the quality of delivered video in the context of immersive video conferencing. Conducting a web-based mean opinion score (MOS) VQA provides more scalability which leads to targeting a wider range of subjects. Moreover, it is cheaper and substantially faster comparing to laboratory quality studies. However, the subjects are not supervised and the equipments and display configuration may vary from one subject to another. In order to avoid any undesired impact of the mentioned issues on the result of the study, the statistically invalid scores were detected and discarded based on the ITU-R BT 500.11 recommendation. Additionally, the web-based system was developed in a way to adapt to the subjects’ environments and provide identical interface based on the subjects’ displays.

This chapter is structured as follows: in Section A.2 the implementation and experimental setting of the system is explained; In Section A.3 and A.4 the subject recruitment and training are investigated; and finally the demographics of the subjects are presented in Section A.6.
A.2 Development of VQA system and experimental setup

Development of the assessment system was consisted of four main sections. The back-end of the system was developed using C#. This part was responsible for the control and logic of the system. Adobe Flash was utilised to display the recorded immersive environment including the avatars with the reference and target video streams on their front surfaces. The user interface of the system was implemented by MVC .net. It was also in charge of maintaining the consistency of the interface on different platforms and presenting the progress bar, the ITU-R ACR scale, Adobe Flash files and other required interfaces. Finally, a SQL server was used to store the collected scores from the subjects.

Since the conducted study was of the web-based type, direct setting of the environment and display configuration was impossible. However, the recorded immersive environment was configured to use a resolution of 790×410 pixels on all platforms and displays. The avatars with the video streams were also positioned at specific distances in each question for all the subjects.

A.3 Subject recruitment

After developing the assessment system, it was extensive tested on multiple platforms with various displays to ensure that the presentation of the user study is consistent and independent of the platforms. Finally, the VQA system was deployed to a server accessible via the internet. Then, electronic participation requests were sent to the students and staffs of the University of Wollongong (UOW) as well as the Smart Services Cooperative Research Centre’s (SSCRC) members and staffs. Moreover, tens of posters were printed and distributed in the university to recruit more subjects. An 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 requested from them, how 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.

Note that the survey and this request was approved by the Human Research and Ethics Committee (HREC) of the university.

A.4 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 where audio was added to the system, the avatar with the audio component also described the instruction of the question to the subject.

Due to the nature of the system, The subject were not tested for vision problems.

A.5 Flow of the study

In addition to the brief description of the user study available on the first page of the system, a button was presented and labelled as ”Go to survey”, which provided an access to the survey. By clicking the button, subjects were redirected to the registration page in which the identity of the subjects was verified. Moreover, information such as their email addresses, genders, age groups and level of educations were 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.
As explained in Section 5.2.3, each subject was free to skip or redo each question with the exception of the last two. If a question was skipped, no score was recorded for that particular question. In the case of redoing 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 score which was closer to the average of all submitted scores to the perceptual scale was maintained 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 it later.

The reason that the last two questions could not be redone was the purpose of the questions. Questions 11 and 12 were analysing the impact of focal point on perceptibility of spatial degradation of video quality and by redoing the questions, the subjects could focus on the avatar, which would not be in their centre of attention area in the first attempt. Hence, the subject were prevented from watching the last two questions multiple times.

After answering the last question, the subject was informed that an email was sent to his/her registered email address to validate that the subject owns the email address. A link was provided in the content of the email to validate the address. Only the validated addresses were entered to the radon draw prize.

A.6 Subjects’ demographics

In this section, subject demographics based on the collected data from the registration page is presented.

A.6.1 Distribution of subject by the last answered question

In addition to the situations (skipping and redoing) mentioned in Section A.5, a subject could leave the survey incomplete. However, the subjects that did not finish the survey were detected and disqualified from entering into the random draw prize to win an Apple iPad 2. As it is demonstrated in Figure A.1,
Experimental Settings and Subjects’ Demographics

Figure A.1 Distribution of subjects by the last answered question

Question 12 is the last answered question by 140 subjects, which also could be interpreted that the majority of the subjects participated in the survey have finished it completely.

Note that discarding invalid scores and disqualification of subjects for the draw prize were completely independent and the scores of uncompleted (e.g. two questions are answered out of 12) but valid surveys are used in the study.

A.6.2 Distribution of subject by gender

One of the collected data in the registration page was the gender of the subjects. The results show that the overall of 118 females and 115 males participated in our user study. As illustrated in Figure A.2, number of female subjects
Experimental Settings and Subjects’ Demographics

Figure A.2 Distribution of subjects by gender

participated in the study is slightly more than male subjects (51% vs. 49%). However, distribution of gender among the subjects is fair and does not show overabundance of women. This slight difference might reflects gender differences in motivation.

A.6.3 Distribution of subject by age

In this section, age distribution of subjects is analysed. In the registration page, nine age groups were provided in a drop down menu to be selected. The age groups were as follows:

- ‘Under 15’
- ‘16-20’
- ‘21-25’
- ‘26-29’
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 among different age group is presented in Figure A.3.
Since, the participation request was mainly sent to the universities and research centres, the subjects found to be highly educated. 184 out 233 subjects participated in the study had education level of Bachelor’s degree or higher. The distribution subject by their education levels is demonstrated in Figure A.4.

**Figure A.4 Distribution of subjects by education**

**A.6.4 Distribution of subject by education**

Since, the participation request was mainly sent to the universities and research centres, the subjects found to be highly educated. 184 out 233 subjects participated in the study had education level of Bachelor’s degree or higher. The distribution subject by their education levels is demonstrated in Figure A.4.
Appendix B

Data Analysis of The User Study

B.1 Introduction

This Appendix documents the collected data of the user study. In Chapter 5, the processed data is presented and it is attempted to demonstrate the perceptual impact of 3D states of avatars on spatial and temporal degradation of video quality, while in this appendix, the data is presented from different angel. These information may provide a chance for other researchers to have their own impression of the conducted study in order to utilise them in their own works.

As explained in Section 5.2.4, five button based on ITU-R ACR scale displayed on the bottom of each question (except last two questions). The wordings are modified to better suit the purpose of the study. Five equally spaced button were displayed as follows: “Identical quality”; “Not identical but hardly noticeable degradation”; “Slightly noticeable degradation”; “Noticeable degradation”; and “Unacceptable degradation”.

In this appendix, the contribution of subjects in each question is presented. It is also tried to demonstrate density of votes to each perceptual scale (button) for every question.
B.2 Subjects’ contribution and votes’ distribution

Since a web-based MOS study is conducted and subjects were free to skip or redo a question, contributions to the questions vary. However, in the processing phase, the statistically invalid scores are discarded and the redundant scores that are submitted by the same subject are filtered. In the filtering mechanism the score which was closer to the average of all submitted scores of that particular perceptual scale was maintained and the other scores submitted by the subject to the same question were discarded.

B.2.1 Virtual distance vs. spatial resolution

The impact of virtual distance on spatial resolution is studied in Question 1 to 3. In all three question, the same spatial degradation is applied (see Section 5.2.5.1). Density of votes to the perceptual scales presents the number of subjects that perceive the chosen quality (label of the button) during the demonstration of that slice in the degraded video sequence.

B.2.1.1 Question 1

In this question, the avatar is located at 9 meters virtual distance to the camera. Since this question was the first question, it had the highest contribution.

As shown in Figure B.3, subjects started detecting a slight noticeable degradation at the resolution of 116×94 pixels and finally 74×60 pixels spatial size at 9 meters distance was recognised as unacceptable degradation to the majority of subjects (Figure B.5).

In Figure B.6 to B.10, mean, median, mode and standard deviation of scores submitted to Question 1 by pressing the corresponding buttons are presented. The figures are plotted in the order of recorded scores. It should be considered that subjects have mainly pressed some of the available buttons (e.g. two or three button out of five buttons) and a minority of participants pressed all
Data Analysis of The User Study

Figure B.1 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 1

Figure B.2 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 1
Figure B.3 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 1

Figure B.4 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 1
available button in a question. Hence, the first score recorded in the first and second perceptual scale are not necessarily submitted by the same subject.

B.2.1.2 Question 2

In this question, the avatars are moved toward the camera and they are located at a distance of 6 meters with respect to the camera’s position. The same spatial degradation procedure is applied on the video of the target avatar. As expected, subjects detected the spatial degradation in higher resolutions. For instance, the resolution of 116×94 pixels was submitted as “Unacceptable degradation” for 6 meters virtual distance, while degradation to this resolution could slightly be noticed at 9 meters distance.

The statistical measurement of distribution of submitted scores for each perceptual scale in Question 2 is demonstrated in Figure B.16 to B.20.
Figure B.6 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 1

Figure B.7 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 1
Figure B.8 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 1

Figure B.9 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 1
Figure B.10 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 1

Figure B.11 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 2
Data Analysis of The User Study

Figure B.12 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 2

Figure B.13 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 2
Figure B.14 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 2

Figure B.15 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 2
Figure B.16 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 2

Figure B.17 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 2
Data Analysis of The User Study

Figure B.18 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 2

Figure B.19 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 2
Data Analysis of The User Study

Figure B.20 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 2

B.2.1.3 Question 3

In Question 3, the avatars are located at the closest distance to the camera (3 meters). The target avatar’s video sequence starts with the slice having the spatial resolution of 352×288 pixels and ends with the slice at the spatial resolution of 48×38 pixels. The temporal resolution is constant and fixed at 20 fps.

Due to the position of the avatars and bigger projection of the video surface on the screen, the video quality drop is more perceivable as demonstrated in Figure B.21 to B.30. It also seems that subjects were trained during the first two questions, as higher number of subjects have voted for same spatial resolution in the same perceptual category.

B.2.1.4 Question 4

The impact of virtual distance on temporal resolution is studied in Question 4 to 6. The spatial resolution in these three questions is fixed at CIF resolution,
Data Analysis of The User Study

Figure B.21 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 3

Figure B.22 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 3
Figure B.23 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 3

Figure B.24 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 3
Data Analysis of The User Study

Figure B.25 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 3

Figure B.26 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 3
Figure B.27 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 3

Figure B.28 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 3
Figure B.29 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 3

Figure B.30 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 3
while the temporal resolution is degraded in 10 step-sizes (See Table 5.1). In Question 4, the avatars are positioned at 9 meters virtual distance.

As the result of locating the avatars at the most distant location, it was expected that degradation of temporal resolution also be harder to perceive. The results also confirm the hypothesis. However, the tolerance of subjects to the low frame rate was higher than the researchers’ expectation. Since Subjects were not aware of the type of degradations and the spatial degradation was assessed in the first three questions, they might have been affected by the expectation of having another spatial degradation. Hence, the temporal degradation could be ignored until it was completely noticeable (very low frame rates).

B.2.1.5 Question 5

In this question, the temporally degraded video is applied on the front surface of the target avatar located at 6 meters away from the camera.
Figure B.32 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 4

Figure B.33 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 4
Figure B.34 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 4.

Figure B.35 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 4.
Figure B.36 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 4

Figure B.37 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 4
Data Analysis of The User Study

**Figure B.38** Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 4

**Figure B.39** Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 4
Figure B.40 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 4

Figure B.41 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 5
Figure B.42 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 5

Figure B.43 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 5
**Figure B.44** Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 5

**Figure B.45** Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 5
Data Analysis of The User Study

As expected, subjects could detect the frame rate drop at the higher temporal resolutions, when the avatar is closer to the viewpoint.

B.2.1.6 Question 6

Question 6 was the last question analysing the impact of virtual distance on the subject’s perception of different temporal resolutions. In this question, the avatars are located at a distance of 3 meters to the camera in order to have the biggest projection of the video surface on the screen.

Unlike spatial resolution, the virtual distance does not have a solid impact on perception of the temporal resolution reduction. It seems that only when avatars are located beyond a certain distance from the camera, the degradation of the temporal resolution is not perceivable to the subjects. Thus, a region-based mechanism is proposed, in which the environment is partitioned to three regions\(^1\). The regions are circular and constructed based on the subjects’ sen-

\(^1\)In this method, the regions are constructed with the local client located at the centre
**Figure B.47** Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 5

**Figure B.48** Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 5
**Figure B.49** Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 5

**Figure B.50** Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 5
sitivity threshold to video frame rate drop beyond the chosen distances. The bandwidth saving achieved using this method is discussed and demonstrated in Section 5.4.

B.2.1.7 Question 7

In the next four questions, the effect of orientation of avatars on perceptibility of video quality degradation is studied. Hence, the avatars’ positions are fixed at 3 meters distance from the camera and only the avatars’ orientations are changed. In Question 7, the avatars are rotated by 30 degrees and the spatial resolution of the target avatar is reduced gradually.

Since 30 degrees rotation is not a huge orientation change and the projected size and shape of the video surface is very close to the original projection (0 of them. The first region is up to 6 meters distance from the camera, any avatar with a distance of 6 to 9 meters is at the second region and beyond 9 meters is considered as the third region. The frame rates of 17, 13 and 10 fps is chosen for the first, second and third region respectively.
Figure B.52 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 6

Figure B.53 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 6
Figure B.54 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 6

Figure B.55 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 6
Data Analysis of The User Study

Figure B.56 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 6

Figure B.57 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 6
Figure B.58 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 6

Figure B.59 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 6
Figure B.60 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 6

Figure B.61 Number of subjects voted for “Identical quality” at different spatial resolutions in Question 7
Figure B.62 Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 7

Figure B.63 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 7
Data Analysis of The User Study

Figure B.64 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 7

Figure B.65 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 7
Data Analysis of The User Study

Figure B.66 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 7

degrees rotation), the outcomes are very close the achieved results in Question 3. However, the change of orientation slightly affected the perceived spatial resolutions and caused the lower resolutions to be less perceivable.

B.2.1.8 Question 8

Question 8 analyses the impact of perceived spatial resolution again. However, in this question, the avatars are rotated by 60 degrees.

Due to the higher rotation of avatars, the effect of the avatars’ orientation is more noticeable. However, the impact of increasing relative angle is not as effective as the impact of virtual distance. The reason is the employed degradation method, as the whole image is uniformly degraded which is suitable for the change of virtual distance but not orientation (see Chapter 6).
Figure B.67 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 7

Figure B.68 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 7
Figure B.69 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 7

Figure B.70 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 7
Data Analysis of The User Study

**Figure B.71** Number of subjects voted for “Identical quality” at different spatial resolutions in Question 8

**Figure B.72** Number of subjects voted for “Not identical but hardly noticeable degradation” at different spatial resolutions in Question 8
Figure B.73 Number of subjects voted for “Slightly noticeable degradation” at different spatial resolutions in Question 8

Figure B.74 Number of subjects voted for “Noticeable degradation” at different spatial resolutions in Question 8
Figure B.75 Number of subjects voted for “Unacceptable degradation” at different spatial resolutions in Question 8

Figure B.76 Mean, median, mode of the submitted spatial resolutions for “Identical quality” in Question 8
Figure B.77 Mean, median, mode of the submitted spatial resolutions for “Not identical but hardly noticeable degradation” in Question 8

Figure B.78 Mean, median, mode of the submitted spatial resolutions for “Slightly noticeable degradation” in Question 8
Figure B.79 Mean, median, mode of the submitted spatial resolutions for “Noticeable degradation” in Question 8

Figure B.80 Mean, median, mode of the submitted spatial resolutions for “Unacceptable degradation” in Question 8
Data Analysis of The User Study

B.2.1.9 Question 9

In Question 9, the avatars are located at a distance of 3 meters and rotated by 30 degrees. The video sequence is degraded temporally in 10 step-sizes according to Table 5.1 and applied on the target avatar’s video surface. The submitted scores by the subject are presented in Figure B.81 to B.90.

The impact of orientation on the perceived temporal degradation, especially for the small rotations such as this question is negligible.

B.2.1.10 Question 10

Although the orientation of the target avatar with respect to the camera is increased through a higher rotation (60 degrees), a minor change is observed in the achieved outcomes.

The results show that the perceived temporal resolution is barely dependant on
Figure B.82 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 9

Figure B.83 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 9
Figure B.84 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 9

Figure B.85 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 9
Data Analysis of The User Study

Figure B.86 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 9

Figure B.87 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 9
Figure B.88 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 9

Figure B.89 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 9
Figure B.90 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 9

Figure B.91 Number of subjects voted for “Identical quality” at different temporal resolutions in Question 10
Figure B.92 Number of subjects voted for “Not identical but hardly noticeable degradation” at different temporal resolutions in Question 10

Figure B.93 Number of subjects voted for “Slightly noticeable degradation” at different temporal resolutions in Question 10
Figure B.94 Number of subjects voted for “Noticeable degradation” at different temporal resolutions in Question 10

Figure B.95 Number of subjects voted for “Unacceptable degradation” at different temporal resolutions in Question 10
Data Analysis of The User Study

Figure B.96 Mean, median, mode of the submitted temporal resolutions for “Identical quality” in Question 10

the orientation of avatars. However, the skewness of the results is changed and indicates, a higher number of participants voting for lower temporal resolutions relatively. The absolute distribution of scores, when the orientation of avatars are increased has a modest change.
Figure B.97 Mean, median, mode of the submitted temporal resolutions for “Not identical but hardly noticeable degradation” in Question 10

Figure B.98 Mean, median, mode of the submitted temporal resolutions for “Slightly noticeable degradation” in Question 10
Data Analysis of The User Study

Figure B.99 Mean, median, mode of the submitted temporal resolutions for “Noticeable degradation” in Question 10

Figure B.100 Mean, median, mode of the submitted temporal resolutions for “Unacceptable degradation” in Question 10