DYNAMIC ADAPTATION OF STREAMED REAL-TIME E-LEARNING VIDEOS OVER THE INTERNET

Aruna Thakur

A thesis submitted in partial fulfilment for the degree of DOCTOR OF PHILOSOPHY

School of Computing, Communication & Electronics
University of Plymouth

May 2007
This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.
ABSTRACT

Dynamic Adaptation of Streamed Real-Time E-learning Videos over the Internet

Aruna Thakur

Even though the e-learning is becoming increasingly popular in the academic environment, the quality of synchronous e-learning video is still substandard and significant work needs to be done to improve it. The improvements have to be brought about taking into considerations both: the network requirements and the psycho-physical aspects of the human visual system.

One of the problems of the synchronous e-learning video is that the head-and-shoulder video of the instructor is mostly transmitted. This video presentation can be made more interesting by transmitting shots from different angles and zooms. Unfortunately, the transmission of such multi-shot videos will increase packet delay, jitter and other artifacts caused by frequent changes of the scenes. To some extent these problems may be reduced by controlled reduction of the quality of video so as to minimise uncontrolled corruption of the stream. Hence, there is a need for controlled streaming of a multi-shot e-learning video in response to the changing availability of the bandwidth, while utilising the available bandwidth to the maximum.

The quality of transmitted video can be improved by removing the redundant background data and utilising the available bandwidth for sending high-resolution foreground information. While a number of schemes exist to identify and remove the background from the foreground, very few studies exist on the identification and separation of the two based on the understanding of the human visual system. Research has been carried out to define foreground and background in the context of e-learning video on the basis of human psychology. The results have been utilised to propose methods for improving the transmission of e-learning videos.

In order to transmit the video sequence efficiently this research proposes the use of Feed-Forward Controllers that dynamically characterise the ongoing scene and adjust the streaming of video based on the availability of the bandwidth. In order to satisfy a number of receivers connected by varied bandwidth links in a heterogeneous environment, the use of Multi-Layer Feed-Forward Controller has been researched. This controller dynamically characterises the complexity (number of Macroblocks per frame) of the ongoing video sequence and combines it with the knowledge of availability of the bandwidth to various receivers to divide the video sequence into layers in an optimal way before transmitting it into network.

The Single-layer Feed-Forward Controller inputs the complexity (Spatial Information and Temporal Information) of the on-going video sequence along with the availability of bandwidth to a receiver and adjusts the resolution and frame rate of individual scenes to transmit the sequence optimised to give the most acceptable perceptual quality within the bandwidth constraints.

The performance of the Feed-Forward Controllers have been evaluated under simulated conditions and have been found to effectively regulate the streaming of real-time e-learning videos in order to provide perceptually improved video quality within the constraints of the available bandwidth.
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors Dr. Bogdan V. Ghita, Dr. Benn M. Lines, Prof. Paul Reynolds, and Prof. Steven M. Furnell.

I am also grateful to my friends – Atanu Nath and Barbara Holvik – for all the help rendered during the course of this work.

Finally, I would like to thank all my family members for their support.
AUTHOR’S DECLARATION

This is to certify that the thesis entitled “DYNAMIC ADAPTATION OF STREAMED REAL-TIME E-LEARNING VIDEOS OVER THE INTERNET” which is being submitted by the author in the partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy in the School of Computing, Communication & Electronics at the University of Plymouth, is a record of original and independent work carried out by the author, except where specifically acknowledged in the text of this thesis. The matter embodied in this report has not been submitted for the award of any other degree.

Aruna Thakur
Aruna Thakur
May 2007
GLOSSARY

**ADSL.** It stands for **Asymmetric Digital Subscriber Line.** ADSL supports data rates of from 1.5 to 9 Mbps when receiving data (known as the *downstream* rate) and from 16 to 640 Kbps when sending data (known as the *upstream* rate).

**Asynchronous Instruction.** E-learning instruction which takes place when the students choose to and interact with the instructors according to their schedules.

**Bottom-up Processing.** When the perception takes place only under the influence of the visual field itself, or changes in stimuli.

**Change Blindness.** The striking failure to see large changes that normally would be noticed easily.

**CODEC.** In telecommunication, *codec* is an abbreviation for Coder-Decoder. It's an analogue-to-digital (A/D) and digital-to-analogue (D/A) converter for translating the signals.

**Fovea.** Region in the centre of the eye, that is most sensitive to light.

**Inattentional Blindness.** The failure of the humans to see unattended items in a scene.

**Luminance Grating.** A laboratory stimulus used to study spatial vision; it is a striped pattern containing alternating light and dark bars, commonly with a sine wave luminance profile.

**Macroblock.** Each macroblock consists of a 16 x 16 block of pixels in a picture. The number of chroma pixels (C_r, C_b) in a macroblock depends upon the type of YUV colour space. A macroblock in 4:2:0 YUV colour space has 6 8x8 blocks. 4 of these blocks are located in the Y plane, 1 in the U/C_b plane, and 1 in the V/C_r plane.

**Spatial Contrast Sensitivity.** The reciprocal of spatial contrast threshold (1/threshold); higher values indicate better performance.

**Spatial Contrast Threshold.** The minimum contrast between the lightest and darkest parts of a pattern required for it to be reliably detected by an observer; lower values indicate better performance.

**Spatial Frequency.** A measure of the fineness of a grating’s bars, in terms of the number of grating cycles (bright-dark bar pairs) per degree of visual angle.

**Spatial Information.** It is calculated according to the ITU-T Recommendation P.910 1999. It gives an indication of the added or lost edges in the destination scene compared to the reference scene. Added edges result from impairments such as tiling, error blocks, and noise. Lost edges can result from impairments such as blurring.

**Spatiotemporal Contrast Sensitivity.** Sensitivity to combinations of spatial modulation and temporal modulation, namely luminance gratings with flicker bars.
**Synchronous Instruction.** E-learning instruction which takes place in real-time, hence requires the simultaneous participation of all students and instructors.

**Temporal Contrast Sensitivity.** The reciprocal of the amount of contrast between the brightest and darkest phases of a flicker stimulus required for a subject to detect the flicker.

**Temporal Frequency.** The alternation rate of a flicker stimulus, measured in cycles/s (Hertz), or the number of flicker cycles (bright-dark alternations) per second.

**Temporal Information.** It is calculated according to the ITU-T Recommendation P.910 1999. It gives an indication of the added or lost motion in the estimation scene compared to the reference scene. Added motion results from impairments such as jerkiness, error blocks and noise. Frame repetition introduces lost motion.

**Top-down Processing.** When the perception takes place under the influence of the expectations, memories, biases and other general upper-level parameters.
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>i</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>AUTHOR'S DECLARATION</td>
<td>iii</td>
</tr>
<tr>
<td>GLOSSARY</td>
<td>iv</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Problems Faced While Using Synchronous E-Learning Videos</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Need for Further Study</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Aim of the Research</td>
<td>5</td>
</tr>
<tr>
<td>1.5 The Scope of Work</td>
<td>5</td>
</tr>
<tr>
<td>1.6 Organization of the Thesis</td>
<td>7</td>
</tr>
<tr>
<td>CHAPTER 2 LITERATURE REVIEW</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Introduction to E-Learning</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Multimedia Streaming over Internet</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Video Transmission over a Best-Effort IP Network</td>
<td>22</td>
</tr>
<tr>
<td>2.4 Video Compression Techniques Based on the Separation of Background and Foreground</td>
<td>36</td>
</tr>
<tr>
<td>2.5 Basics of the Human Visual System</td>
<td>40</td>
</tr>
<tr>
<td>2.6 Degradation of Video During Transmission</td>
<td>49</td>
</tr>
<tr>
<td>2.7 Perceptual Quality Measurement</td>
<td>59</td>
</tr>
</tbody>
</table>
CONTENTS

CHAPTER 3 PROPOSED ARCHITECTURE 70

3.1 Introduction 70
3.2 Requirements for the Transmission of an E-learning Video over the Internet 70
3.3 Set-Up for a Virtual Class Room 72
3.4 Optimum Transmission of Video 74
3.5 Separation of Background and Foreground 75
3.6 Adaptation of Video Transmission Using Feed-Forward Controllers 77
3.7 Adaptation of Video Transmission Using Multi-Layer Feed-Forward Controllers 79
3.8 Adaptation of Video Transmission Using Multi-Layer Feed-Forward Controllers 83
3.9 Codecs Used for the Research 85
3.10 Preparation of Test Video Sequences 86
3.11 Characterisation of Video Sequences 88
3.12 Summary and Conclusions 93

CHAPTER 4 HUMAN VISUAL SYSTEM BASED FOREGROUND-BACKGROUND SEPARATION 95

4.1 Introduction 95
4.2 Overview of the Psychophysical Experiments 96
4.3 Experimental Procedure 97
4.4 Experiment 1 – Identification of the Objects of Central and Marginal Interests 99
4.5 Experiment 2 – Relationship Between the Inattentional Blindness and the Perceptual Quality of Video 103
4.6 Experiment 3 – Relationship Between the Change Blindness and the Perceptual Quality of Video 108
4.7 Discussion on the Results of Psychological Experiments 113
4.8 Separation of Foreground 114
4.9 Characterisation of Background 115
4.10 Separation of Instructor from the Frames 118
4.11 Separation of Whiteboards from the Frames 127
4.12 Addition of Foreground over Background 130
4.13 Summary and Conclusions 132
CONTENTS

CHAPTER 5 OPTIMISED TRANSMISSION OF VIDEO IN LAYERS USING MULTI-LAYER FEED-FORWARD CONTROLLER
5.1 Introduction 134
5.2 Functional Architecture of the Multi-Layer Feed-Forward Controller 134
5.3 Dynamic Characterisation of the Ongoing Video Sequence 136
5.4 Prediction of the Quality of Incoming Video Frames 137
5.5 Development of the Controller Equations 138
5.6 Calculation for the Selection of the Best Layering Technique 150
5.7 An Example of the Calculation for the Selection of the Best Layering Technique 151
5.8 Evaluation of the Performance of the Multi-Layer Feed-Forward Controller 153
5.9 Summary and Conclusions 158

CHAPTER 6 DESIGN OF SINGLE-LAYER FEED-FORWARD CONTROLLER
6.1 Introduction 159
6.2 Functional Architecture of the Single-Layer Feed-Forward Controller 159
6.3 Dynamic Characterisation of the Ongoing Video Sequence 161
6.4 Prediction of the Quality of Incoming Video Frames 165
6.5 Development of the Controller Equations 169
6.6 Calculation of the Available Options 176
6.7 Selection of the Best Option from the Available Options 177
6.8 Summary and Conclusions 179
CHAPTER 1

INTRODUCTION
CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

The Instructional Technology Council\(^1\) [2006] defines distance education as "the process of extending learning, or delivering instructional resource-sharing opportunities, to locations away from a classroom, building or site, to another classroom, building or site by using video, audio, computer, multimedia communications, or some combination of these with other traditional delivery methods". It is a cost-effective way of providing education to students in their own home or office and is especially suited to students who are located in geographically distant places and who cannot physically attend classes due to various reasons. It is also becoming popular among people who are working and are in need of the further development of their skills but cannot take out time to attend regular classes [CDLP 2006].

With the advancements in Internet technology, the style of distance education has undergone a major change. Nowadays distance education involves extensive use of the Internet for transmitting lectures and for interaction between students and teachers. These distance education classes, henceforth called e-learning classes, can be classified into two different categories, based upon the interaction between the instructor and students; these are asynchronous and synchronous [CDLP 2006].

- In *asynchronous e-learning*, educational materials are stored on a server and students retrieve them at a time convenient to themselves.

- In *synchronous e-learning* the instruction involves real-time delivery of multimedia materials. In this scenario the teacher and students interact in real time.

---

\(^1\) The Instructional Technology Council (ITC) represents higher education institutions in the United States and Canada that use distance learning technologies.
Due to the relative ease in organising and presenting the instruction material, asynchronous methods have become quite common, but with the improvements in the performance of videoconferencing tools, more and more colleges are experimenting with synchronous methods. In future, some hybrid methods, involving both synchronous and asynchronous methods, may become more common. It is quite possible that a virtual classroom could become very popular so that the students who can attend the class in person will, while those who cannot will take part by remote connection [Sigle 2001].

1.2 PROBLEMS FACED WHILE USING SYNCHRONOUS E-LEARNING VIDEOS

In spite of its growing popularity, the use of video in synchronous e-learning suffers from a number of problems.

- The courses are ill planned and ill presented since many colleges cannot afford state-of-the-art equipment and the teachers are not properly trained [Hentea et al. 2003, Zhang et al. 2004].

- Students only see a head-and-shoulder presentation of their instructor and often with poor quality video, so even if the course material is well-presented the instructor is handicapped because he is forced to present it while seated in front of the camera [Hentea et al. 2003).

- In a real classroom, the body language of instructor and the students, like facial expressions, eye gaze and gestures, play an important part in nonverbal communication and they direct the attention of students to the relevant area. Due to the limitations of the head-and-shoulder presentation and the poor video quality, video often fails to convey these clues; as a result, instructors are unable to make the class attractive and effectively guide the students; and students are unable to give clear feedback to the instructor [Atkinson et al. 2002, Hentea et al. 2003, Zhang et al. 2004].

- The available bandwidth in the network changes dynamically and if the sender does not adjust its transmission rate, etc., based on the available bandwidth, it may result in random packet drops, jitter, etc. This may result in poor video quality at the students’ end [Crowcroft et al. 1998].
- Even if the instructor does manage to involve students and gets them to carry out their tasks satisfactorily, the poor video quality forces the students to expend extra effort in understanding and deciphering the information. This leads to conscious/unconscious discomfort or distress. This distress manifests itself in the form of psycho-physiological disorders like mental stress, higher heart beat rate and overall tiredness. This is damaging to the subjects in the long run [Atkinson et al. 2002, Wilson 2001].

These deficiencies can result in no, or at best, limited interaction among participants, resulting in boredom and disengagement. To overcome these problems two important issues need to be resolved:

- **Presentation Issues.** The instruction video should be prepared, keeping in mind the principles of cinematography so as to excite the interest of students.

- **Technical Issues.** This involves capturing and transmission of video and audio in a quality comparable to TV quality.

Over the years the art of movie making has evolved so as to make film an effective media of communication. These techniques have also been adopted with certain modifications for “small screen” presentations like TV. Today, even simple events like group discussions, newscast, speeches, etc. can be presented in an interesting and effective way; but creating such an interesting and effective video is an art that requires considerable talent and training [Gleicher et al. 2002].

Unlike the real classroom, where the students take part in the lectures, in e-learning the control is left primarily in the hands of the instructor. Surveys have shown that students interact with instructor and themselves far less in e-learning courses than in real classes [Chen 2001]. Hence, it becomes all the more important for the instructor to efficiently present the lecture material and guide the students through his or her presentation without putting too much burden of comprehension on the students. This is possible only if the instructor is allowed the freedom to present the lecture in his or her own style, i.e., not being restricted by the head-and-shoulder video, and if the video is transmitted in good quality [Gleicher et al. 2002].
1.3 NEED FOR FURTHER STUDY

A considerable amount of work has been done in improving the quality of real-time video transmission. On the hardware side the processing power of the computers and the availability of bandwidth have increased manifold. On the software side, a significant amount of work has been done in improving the codecs and transmission protocols. In spite of the tremendous improvements in the last couple of decades, a lot of work needs to be done to make the e-learning videos acceptable to the students.

While it is possible to produce an exciting video for instructional purposes, employing a number of skilled production crew in a studio, it is not economically viable for most colleges on a daily basis when they want to deliver lower-cost education via e-learning. Hence, there is a need for the production of cost-effective video following the basic principles of filmmaking. While it may not be possible to automate the “art” of filmmaking; within a limited framework it should be possible to video the classroom lectures in a more interesting manner [Gleicher et al. 2002].

One simple way to make the lecture video more interesting than a talking head-and-shoulder video is to use a variety of shots taken from different angles and views [Videomaker Inc. 2004]. While the use of various scenes will improve the presentation quality of video, it will also present a number of technical challenges. One of the challenges that will arise will be the transmission of video without intolerable packet drops, jitters and other artifacts. While these problems are common to any video transmission, they could become more severe for video involving frequent changes in the scenes because different scenes have different bandwidth requirements. In the best-effort IP networks, which are unreliable and unpredictable, the sudden changes in scenes may result in uncontrolled packet drops and delays which may affect the smooth play-out of compressed video. This is because noise generated by packet loss or delay may get propagated due to the inter-frame dependencies.

The severity of video degradation caused by the packet losses and delays can at times be reduced by controlled reduction of the video quality. This is because an uncorrupted video stream having slightly lower quality is less irritating to the user than a corrupted high quality stream. However, rapidly fluctuating quality should also be avoided since the human vision system adapts to a specific quality after a few seconds, and it becomes annoying if the viewer has to continuously adjust to a varying quality. Therefore there is a need for transmitting multi-scene e-learning video in a controlled manner to overcome the problems caused by the frequent changes in the video scene whilst optimally utilising the available bandwidth.
1.4 AIM OF THE RESEARCH

Aim of the study is to develop an understanding for the adaptation of the multi-scene video streaming for e-learning purposes taking into consideration the available bandwidth and the user-perceived quality over best-effort IP networks.

Hence, the goal of this research work is to gain an insight into the human psycho-physical visual system and then use the understanding for the adaptation of multi-scene video transmission for e-learning purposes over the Internet. The focus of the adaptation of the video has been to reduce the amount of information to be transmitted and then to stream the video with the intention to maximise the user-perceived quality.

1.5 THE SCOPE OF WORK

This research proposes that it is possible to maximise the perceptual quality of multi-scene video for e-learning applications by adapting the transmission in response to the changing bandwidth availability. This can be done by first removing the redundant background information and then by adjusting the video transmission parameters so as to maximise the perceptual quality at the receiver end.

A number of schemes exist for identification between foreground and background, however very few studies have exploited the understanding of the psycho-physiological aspects of human visual for this task. As a first step it is proposed that the foreground and background be defined in the context of the e-learning videos. It is then proposed that the differing roles inattentional blindness\(^2\) and change blindness\(^3\) on foreground and background be understood for gaining a greater insight into the characteristics of the two portions of video sequences. Once the two portions in a sequence of frames have been characterised, they may be separated by suitable techniques.

---

\(^2\) Inattentional blindness is inability to perceive something because of diversion of attention by other task.

\(^3\) Change blindness is the inability to note large changes in two similar photos or video sequences.
The research further proposes that having separated background section from the sequence of frames, the video may then be efficiently transmitted within the available bandwidth in the perceptually optimal way. In order to dynamically adapt the video transmission a concept of **Feed Forward Controller (FFC)** has been proposed. This controller dynamically characterises the ongoing video sequence in terms of the complexity, predicts the characteristics of the incoming sequence and then adapts the transmission so as to stream the video within the available bandwidth. Such dynamic adaptation intends to reduce the chances of uncontrolled packet drops, thereby improving the perceptual quality at the receiver end.

The first controller that has been researched, **Multi-Layer Feed-Forward Controller**, is a post-codec controller that provides a service to users connected with various access bandwidths so as to ensure intra-session fairness, by which all the users receive data at a rate appropriate to their own capabilities regardless of the capacities of other users. This basic requirement poses major problem because of the intrinsic heterogeneity and large scale of the Internet. In order to provide scalability, the controller dynamically characterises the complexity of the ongoing video sequence in term of number of Macroblocks per frame and combines the knowledge of the bandwidth availability to various receivers to partition the video sequence into layers in an optimal way before transmitting it into network. The controller adjusts the number of layers, distribution of frames among the layers and controls the frame rate in order to provide perceptually optimised video to a group of receivers connected by varied bandwidth links in a heterogeneous environment.

The second controller that has been studied, **Single-Layer Feed-Forward Controller**, is an add-on module to a codec. The controller dynamically characterises the ongoing video sequence in terms of Spatial Information and Temporal Information and inputs the availability of bandwidth to a receiver. Based on these three values, it then dynamically adjusts frame rate and resolution of the streamed video sequence. These adaptation parameters have been selected because they most closely correspond to the complexity – Spatial Information (spatial details) and Temporal Information (sequential changes) – of the video sequences. The controller thus dynamically adjusts its transmission in order to effectively utilise the available bandwidth in order to provide perceptually best possible video.
Hence, the current research focuses on the following:

- Defining the foreground and background in the context of e-learning video.
- Understanding the roles of inattentional blindness and change blindness on foreground and background portions of the video sequences.
- Regulated transmission of layered video to provide perceptually optimised video to a group of receivers in a heterogeneous environment.
- Regulated transmission of video within available bandwidth in the perceptually optimal way.

1.6 ORGANISATION OF THE THESIS

Chapter 2 of this thesis is devoted to a critical review of literature on the relevant aspects. In this chapter the fundamentals of the video transmission over the Internet have been described. The running of multimedia applications over the Internet not only gives poor results in terms of applications but also affects the performance of other data applications. The cause of poor performance and the possible solution for them have been discussed. One of the ways of improving the quality of video transmission over Internet is the dynamic adaptation of video quality. The chapter describes various schemes that have been considered in terms of their strengths and weaknesses.

The requirements for the optimal transmission of e-learning video sequences have been described in Chapter 3. The chapter also describes the proposed architecture to transmit the video within the available bandwidth so as to provide the best perceptual quality to the viewers.

The definitions of “foreground” and “background” in the context of e-learning videos have been described in Chapter 4. That chapter further describes the roles of “inattentional blindness” and “change blindness” on the perception of the two portions of the video. Finally, the chapter describes the work carried out to separate foreground and background.
Chapter 5 describes the post-codec Multi-Layer Feed-Forward Controller that carries out temporal layering of the video stream before transmission. This controller dynamically characterises the on-going video sequence in terms of number of Macroblocks per frame, inputs the available bandwidth from the receivers and then partitions the video stream into layers in an optimal way before transmitting it into network. The performance of this controller has also been discussed in this chapter.

Chapter 6 describes the Single-Layer Feed-Forward Controller that considers bandwidth requirement and perceptual quality of video as dependent variables and the complexity of video sequence (Spatial Information and Temporal Information), frame rate and resolution as independent variables. The controller uses the relationships between: (a) Frame rate – Resolution – Bitrate; and (b) Spatial Information – Temporal Information – Perceptual Quality – Bitrate in order to recommend the best transmission condition (frame rate and resolution) for a video sequence.

The performance of the Single-Layer Feed-Forward Controller has been described in Chapter 7. The performance of the architecture has been evaluated in a simulated condition and the results obtained have been discussed in the chapter.

Major conclusions from the present study and suggestions for future work are given in Chapter 8. Relevant references are included at the end of the thesis.
CHAPTER 2

LITERATURE REVIEW
CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION TO E-LEARNING

The California Distance Learning Project (CDLP) [2006] defines distance learning as follows: “Distance Learning (DL) is an instructional delivery system that connects learners with educational resources. DL provides educational access to learners not enrolled in educational institutions and can augment the learning opportunities of current students. The implementation of DL is a process that uses available resources and will evolve to incorporate emerging technologies.”

The distance education delivery system can be of two types:

- **Synchronous instruction** which takes place in real-time, hence requires the simultaneous participation of all students and instructors. The advantage of synchronous instruction is that it allows live interaction between the instructors and students. Examples include interactive telecourses, teleconferencing and web conferencing, and Internet chats [CDLP 2006].

- **Asynchronous instruction** which takes place when the students choose to and interact with the instructors according to their schedules. Hence, asynchronous instruction does not require the simultaneous participation of all students and instructors. Examples of asynchronous delivery include e-mail, audiocassette courses, videotaped courses, correspondence courses and WWW-based courses [CDLP 2006].

Figure 2.1 summarises the most popular California distance learning approaches in 2005-2006. The figure shows that the video checkout is still the most popular [CDLP 2006].

Online education refers to any form of learning / teaching that takes place via a computer network. This is also often referred to as e-learning. Online education also involves access to databases in the form of text files or multimedia web pages, as well as the exchange of information (e.g., assignments, course materials) via file transfers. With the developments in the Internet network and videoconferencing softwares, web conferencing is becoming quite common [CDLP 2006].
While e-learning is still not the most popular means of instruction, its popularity is growing quite rapidly. It has gained popularity particularly for two types of adult education programmes – professional development and anytime student instruction [CDLP 2006]. It is being increasingly adopted by many schools as a strategic necessity, hence, many schools have either already started or have allocated funds for the necessary infrastructure for Internet and advanced telecommunications network in the schools [Erickson and Siau 2003].

Since e-learning is still in an evolutionary stage and a number of uncertain issues need to be investigated before its impact in the society can be quantified. One of the biggest questions that needs to be answered is the effectiveness of e-learning education in comparison to the traditional classroom education. A number of studies have been carried out to find which of the two is more effective in imparting the education, but a definite conclusion has not as yet been drawn. This could be because there is no easy way of comparing the performance of the two [Simonson 1999, Zhang et al. 2004].

**FIGURE 2.1.** A survey of various instructional media used for the distance education in California in 2005-2006 [CDLP 2006].
The growth of distance learning has been attributed to a wide range of advantages it offers to the students and educators. Students find it attractive because they are not forced to attend the local college, which may not have courses or facilities meeting their needs. They also like the convenience of learning at their own time and pace. Educators appreciate the opportunities for interactivity offered by e-mail, discussion boards, and online resource centres and laud the possibilities for customising and personalising learning as well as for broadening the curricula [Hentea et al. 2003].

Apart from a number of social problems, a major complaint made by students is that of lack of interaction between the instructor and students, which is vital for effective learning. This lack of interaction leads to the feeling of isolation among the students. Due to the physical distances the students are unable to communicate effectively with instructors and other students. In a classroom a significant portion of communication takes place non-verbally by body-language and facial expressions. In e-learning often these are lost; as a result, both – instructors and students – are unable to convey their feelings. Hence, the students feel isolated, confused and often lost, and the instructors are unable to get feedback from the students. On the other hand, increased interaction leads to increased motivation, a higher quality learning experience and overall higher levels of satisfaction [Kies and Williges 2000, Horn 2001, Atkinson et al. 2002, Hentea et al. 2003, Zhang et al. 2004].

2.1.a Interaction and Use of Video in E-learning

As discussed in the previous section, one of the major problems of e-learning is the lack of interaction, as a result of which the instructors and students are unable to communicate effectively using body-language and facial expressions [Tang and Isaacs 1999]. Hence, for the e-learning to be more accepted the technology should incorporate strategies for increasing the instructor-student interaction so as to give a “feel” of a traditional classroom.

It is thus important that the online teaching methodologies involve interactive components for an active learning process. It is generally accepted that learners retain only 10% of what they read and 90% of what they say and do (Dale's Cone of Learning Experience, Figure 2.2). Unfortunately, many online learning classes do not provide a high enough degree of interaction to motivate the learners or enhance the process of retention, as a result of which, learners frequently describe their online learning experience as less than satisfying [Atkinson et al. 2002].
Realising the importance of interaction, present online distance education strives to provide a rich, near-classroom experience to non-classroom students. This is true for both, asynchronous and synchronous, distance education techniques. With the developments in multimedia transmission over the Internet, the use of videos is being exploited to provide a rich, near-"in person" experience to students at remote locations. While video provides one method of increasing interaction and providing the near-classroom experience it alone cannot fully re-create the face-to-face experience of the traditional classroom [Leonard et al. 2003].

There are three approaches for incorporating video over IP in a classroom. These are [Leonard et al. 2003]:

- **Live streaming video** is a real-time transmission of video used for one-to-one or one-to-many delivery. In many ways live streaming video is similar to television broadcasts in that the instructor lectures, but cannot hear, see, or interact with the students.

- **Interactive video** is also a one-to-one or one-to-many transmission of real-time video, but it provides interaction between the instructor and students.

- **Video-on-demand** is one-to-one or one-to-many transmission of pre-recorded and stored lectures, movies, tutorials, etc. They are downloaded at an individual’s convenience.
In all three cases, the quality of video plays a very important role in determining the success of the e-learning session. The students expect the video to be of good quality because poor quality video forces them to expend extra effort to view them. A poor-quality video causes discomfort and stress even if the students are still capable of performing the main task. It has been observed that under the poor viewing conditions the Galvanic Skin Response (GSR), Heart Rate (HR) increases and the Blood Volume Pulse (BVP) decreases. It has even been observed that when the fall in subjective quality of video is not noticed due to lower frame rate, the physiological responses to low frame rate still take place [Wilson 2001]. It has also been observed that very poor quality of video can also prevent full attention to the auditory channel [Kies and Williges 2000].

2.1. b Requirements of a Good E-learning Video

Over the last hundred years the art of movie making has evolved to make film an effective media of communication. These techniques have also been adopted, with certain modifications, for “small screen” presentations like TV. Today, even simple events like group discussions, newscast, speeches, etc. can be presented in an interesting and effective way; but creating such an interesting and effective video is an art that requires considerable talent and training [Gleicher et al. 2002].

Unlike the real classroom, where the students take part in the lectures, in e-learning the control is left primarily in the hands of the instructor. Surveys have shown that students interact with the instructor and themselves far less in e-learning courses than in real classes [Chen 2001]. Hence, it becomes all the more important for the instructor to efficiently present the lecture material and guide the students through his presentation without putting too much burden of comprehension on the students. Unfortunately, not many instructors are trained to be good presenters and video-graphers. Most often the instructor just sits in front of a monitor with a camera placed over it and delivers the whole lecture without any movement. Unfortunately this mere capturing of talking-head-and-shoulder video does not make an interesting video. What is more interesting than this is a “real lecture” video, where the lecturer has the freedom to move around, use electronic whiteboard and show real models as he would do in a real classroom to real audiences. A number of methods can be adopted to break the monotony of a lecture transmission, these include (a) the use of a variety of shots from different angles and views; and (b) maintaining the length of each shot between 6 and 20 seconds [Videomaker Inc. 2004].
FIGURE 2.3. Screen shots of some commercially available videoconferencing softwares.
A survey was carried out in order to see how many commercially available e-learning / video conferencing tools and university e-learning tools satisfy the criteria for effective video displays. Figure 2.3 shows screen shots of some commercially available e-learning / videoconferencing tools. As can be seen, except for MB@learning, all other commercial tools only display head-and-shoulder videos of the participants and there does not seem to be any provision of taking video shots from different angles and zooms. Also in most of the cases the video windows are quite small.

Figure 2.4 shows that only the SMA tool has a provision for displaying video of the instructor from a distance. Similarly the asynchronous e-learning courses offered by the Stanford Centre for Professional Development (not included in the figure) have videos that cover the class from different angles and zooms.
Since, many of the small colleges and universities use commercially available softwares for their e-learning classes; and many university developed video-conferencing tools do not satisfy the requirements of an effective video it may be expected that the students would get sub-optimal videos of the classes.

2.1.c Challenges in Streaming of Real-Time E-Learning Videos

The survey of commercial and university developed e-learning conferencing tools have shown that most of the tools only transmit small head-and-shoulder video of the instructor or participants. There could be various reasons for doing this. Some of them are:

- **The heterogeneous nature of the network.** The users are connected to the network by varied capacities in terms of the computational power and available bandwidth. Since the real-time transmission can only be done at a fixed rate, it is often difficult to satisfy conflicting requirements of the users connected by low bandwidth and those connected by large bandwidth. Hence, most often the transmission is done at the bottleneck rate in order to satisfy all the users (Discussed in **Section 2.3**).

- **Affordability.** While setting a desk-top conferencing system is relatively cheap, setting a complete room-based system could be quite expensive, because room-based system requires expensive computers, tracking cameras, lighting, audio system, etc. [Sigle 2001].

2.1.d Summary of the Literature Review for this Section

From the literature survey carried out it is apparent that:

1. Recent years have witnessed a constant rise in the use of asynchronous and synchronous e-learning courses.

2. In spite of the growing popularity, the students taking e-learning courses complain about lack of interaction in the classes.

3. One of the ways of increasing interaction is by effectively using videos in the e-learning tools.

4. The quality of videos transmitted in the e-learning classes is poor due to technical limitations, e.g. availability of the bandwidth.

5. There is scope for improving the quality of video transmitted in the e-learning classes.
Chapter 2

The brief literature survey has shown that there is a need for improving the quality of e-learning video in order to increase the level of interaction in class and to increase the overall satisfaction of the students. The next section briefly discusses the transmission of video over the Internet.

2.2. MULTIMEDIA STREAMING OVER INTERNET

Multimedia is the combination of text, graphics, drawings, still images, natural and animated video, audio, etc. On the whole it is a complex interaction between the representation, storage, transmission and processing of the digital information. A typical multimedia set-up consists of a server, which encodes an audio-video object and then transmits it via the Internet, and the receiver(s) which decode the object before playing. This is a technology that has the possibility of revolutionising the field of education. It can provide a very powerful tool for enabling and enhancing course delivery for online students [Braun 1998, Leonard et al. 2003].

2.2.a Codecs

The video is initially captured using a camera or any other media player and then sent to the codec (compressor / decompressor). The codec compresses the amount of information to be sent into the network. For example, transmission of an uncompressed NTSC video would require:

\[
\frac{486 \times 720 \text{ pixels}}{\text{frame}} \times \frac{30 \text{ frames}}{s} \times \frac{24 \text{ bits}}{\text{pixel}} = 250 \text{ Mbits/s}
\]  

(2.1)

If such large raw video data is sent to the network, it would heavily congest the network, resulting in random packet drops and delays. The compressed video information then flows from the source to the receiver(s) through the network. At the receiver end, the codec decompresses the data before playing. Codecs can be implemented in software, hardware, or a combination of both. There are various codecs available and all of these have different characteristics [Keller 1998].

Hence, one of the most important parts of the system is the codec. There are a number of codecs which have been developed and standardised. The standardisation has been done by two standardisation bodies: the International Telecommunications Union (ITU) [2006] and the Motion Pictures Experts Group (MPEG) [2006].
2.2.b Different Types of Codecs Used in this Research

**MPEG-4**

The Motion Picture Experts Group (MPEG) was established in 1988 and works to generate the specifications under the International Organisation for Standardisation (ISO) and the International Electro-technical Commission (IEC). MPEG-4 was designed for very low bit-rates to very high bit-rates for applications running on wireless to broadband connections. It is based on object-based compression, which allows individual objects within a scene to be tracked separately and compressed together resulting in very efficient and scalable compression [Moving Picture Experts Group 2006].

**H.261**

The International Telecommunications Union (ITU), which has its headquarter in Geneva, Switzerland is an international organization within the United Nations System. The ITU works with governments and the private sector to coordinate global telecom networks and services. It has developed a number of standards for video codecs. The H.261 video coding standard has been designed by the ITU for video-telephony and video conferencing applications; it is designed for low bit rate transmission (64 to 2048 kbits/sec) with low coding delay [ITU 2006, Keller 1998].

2.2.c Introduction to MPEG-4

MPEG-4 has proved to be a successful codec and has been widely accepted for video-conferencing applications. This is because of its ability to handle very low to very high bit rates [MPEG 2006].

The work on MPEG-4 was started in July, 1993 and the first version, MPEG-4 Version 1, became an international standard in February, 1999. MPEG-4 Version 2 was standardised in early 2000, and work is still continuing on newer versions. Initially, the aim of MPEG-4 was to support only low bitrates, but with modifications in the standard, it can now support a wide range of bitrates. Today, MPEG-4 Video is optimised for: (a) low (<64 kbps), (b) intermediate (64-384kbps), and (c) high (384kbps-4Mbps) bitrates [Moving Picture Experts Group 2006, Puri and Chen 2000].

It can support both constant bitrate (CBR) and variable bitrate (VBR). Due to its wide range of bitrate and flexible philosophy, its targeted applications include [Moving Picture Experts Group 2006]:
Internet Multimedia
Interpersonal Communications (Video-conferencing, Videophone etc.)
Multimedia Mailing
Remote Video Surveillance
Wireless Multimedia
Broadcasting Applications

Since these applications work on different formats, MPEG-4 supports a wide range of formats.

MPEG-4 is based on object-based encoding, where each scene is decomposed into separate video objects. An example of object-based encoding is a news broadcast, where the news person is encoded as a separate foreground Video Object (VO) while the background images compose another object. Video Object motion is achieved by a progression of video object planes (VOPs) [MPEG 2006, Puri and Chen 2000].

The MPEG-4 video supports the representation of video objects of natural or synthetic origin, coding them as separate entities in the bitstream, which the user can access and manipulate (cut, paste, scale, etc.) [MPEG 2006, Puri and Chen 2000].

In the MPEG-4 context, a VO can still be the traditional case of a sequence of rectangular frames formed by pixels; it can also correspond to a sequence of arbitrarily shaped sets of pixels possibly with a semantic meaning, given that this higher level information is somehow made available (e.g. by providing shape or transparency information) [MPEG 2006, Puri and Chen 2000].

The way VOs are identified is not within the scope of the MPEG-4 standard - it is considered as a pre-processing step. MPEG-4 wants to provide the means to represent any composition of objects, whatever the methods used to achieve the composition information. The arbitrarily shaped VOs can be obtained by a variety of means, such as: automatic, or assisted segmentation of natural data, chroma key techniques, or synthetic computer generated data. The video test material currently used in MPEG-4 contains both rectangular and pre-segmented VOs) [MPEG 2006, Puri and Chen 2000].

MPEG-4 Video bitstream syntax consists of a hierarchy of classes: Video Session (VS), Video Object (VO), Video Object Layer (VOL), and Video Object Plane (VOP). The VS class is the highest entity in the class hierarchy and may contain one or more VO (VO, VO2, ...,
VOs), while each VO can consist of one or more layers (VOL₁, VOL₂, ..., VOLₙ) which can be used to enhance the temporal or spatial resolution of a VO. Thus the VOL class is used to support temporal and spatial scalabilities. An instance of a VOL at a given time instant is called a Video Object Plane (VOP) (Figures 2.5 and 2.6). MPEG-4 consists of three types of VOPs [Puri and Chen 2000, Balk et al. 2004]:

- **Intra-coded VOPs (I-VOPs):** They are encoded independently and can be considered as “key” VOPs.
- **Predicted VOPs (P-VOPs):** contain predicted motion data and information about the errors in the predicted values.
- **Bi-directional predicted (B-VOPs):** This VOP depends on the previous and next VOPs.

The VOPs can be sent on different layers. The base layer can have I (Intra) and P (Predicted) frames; the enhancement layer can have B (Bi-directional frame). The base layer provides the basic quality, whereas the enhancement layers provide refinement to the basic quality [Puri and Chen 2000].

**FIGURE 2.5.** Multiplexing of video object planes at the encoder.

**FIGURE 2.6.** De-Multiplexing of video object planes at the decoder.
Another interesting feature of MPEG-4 is the use of sprite. A static sprite is a large still image, describing panoramic background. In MPEG-4 this background view can be synthesised using a set of background images joined together. This non-changing background has to be transmitted only once. Later the foreground can be separated from the image, using an appropriate segmentation technique and transmitted to the receiver. At the receiver end the foreground can be combined with the appropriate portion of the background to form the complete image [Puri and Chen 2000]. Figure 2.7 shows an example of this. This method of transmission in layers can result in considerable saving of bandwidth.

2.2.d Summary of the Literature Review for this Section

From the literature survey carried out it is apparent that:

1. MPEG-4 is a suitable codec for the transmission of e-learning video over IP because it is designed for very low bit-rate to very high bit-rate applications

2. MPEG-4 supports transmission of video in a number of video object planes (VOPs), hence it is possible to transmit important sections of a video in one layer and less important sections in another.

3. Since the transmission of video is a best-effort service, the transmission rate is modified by the sender at the request of the receiver so as to satisfy the existing availability of the bandwidth.
2.3 VIDEO TRANSMISSION OVER A BEST-EFFORT IP NETWORK

2.3.a Problems of Streaming in Multicast Environment

The transmission of video over IP networks poses a number of challenges. This is because [Crowcroft 1998, Wu et al. 2001, Reibman et al. 2004]:

- Video transmission requires high bandwidth for sending a sequence of high quality images.
- The quality of video is normally sensitive to dropped packets. Retransmission of the lost packets is not a good solution since it may lead to feedback implosion and large-scale congestion. Another problem associated with the retransmission of packets is that they may reach the receivers after the time when they are required has passed.
- They do not back off in case of congestion, thus leaving very little bandwidth for data applications to operate.
- The real-time applications in general are less tolerant to packet delays and losses and the video packets sent by the sender may never be received by the client, causing artifacts or frozen frames.
- The real-time video is transmitted at a uniform rate to all the receivers in the network, hence, the source must run at the bottleneck rate, resulting in high-capacity links getting under-utilised; otherwise, it will overload portions of the network having low bandwidth capacity.
- At times the client buffer may starve, resulting in either video that appears jerky or long re-buffering intervals.

2.3.b Rate Shaping in Video Transmission

In MPEG-4 there are three types of coded pictures [Keller 1998, Zhao 2001, MPEG 2006]:

- **Intra (I) Frames.** These are encoded without any temporal prediction. These frames are coded by encoding blocks of pixels using Discrete Cosine Transformation (DCT), quantiser and reordering. The coded blocks are then grouped together in macroblocks.

- **Predicted (P) Frames.** These are encoded using motion prediction from the previous I or P frames. The Y, Cr and Cb components of each macroblock is matched with the neighbouring blocks in the previous frame. The predicted error and the motion vector are encoded and transmitted.
• **B (Bi-directionally Predicted) Frames.** These are encoded using interpolated motion prediction between the previous I or P picture and the next I or P picture in the sequence. Each macroblock is compared with the neighbouring area in the previous and the next I or P picture. The best matching macroblock or the average of the two macroblocks is chosen and the predicted error and motion vector are encoded and transmitted.

The three picture classes are grouped together to form Group of Pictures (GOPs), thus each GOP consists of one I frame followed by P and B frames. The I frame has lowest compression efficiency, P frame has higher compression efficiency and the B frame has the highest compression efficiency [Kuo 1998, Zhao 2001]. Figure 2.8 shows the bitrate required for transmitting a sequence. The sequence was encoded using MPEG-4 that encoded only I and P frames. The figure clearly shows that the I frames require significantly larger bitrates compared to the P frames. The figure also shows the edges of two neighbouring frames detected using Robert's edge detector (See Appendix 1). The figure shows that the I frame is better formed compared to the adjacent P frame.

Since MPEG-4 encodes a video sequence in the form of GOPs consisting of I, P and B frames, the video sequence shows a rapid fluctuation in bitrate due to the effects of scene complexity and motion. The transmission of a sequence encoded with such wild fluctuations is not desirable either for the Constant Bitrate (CBR) or Variable Bitrate (VBR) transmission. Hence, a video server must adjust the transmission rate so as to maximise the signal quality at the receiver [Bolot and Turletti 1998, Zhao 2001].

From the previous sections it is clear that the availability of bandwidth is an important issue since it decides the quality of video transmission. Unfortunately, in the current network the bandwidth availability varies dynamically and one of the ways to solve these problems is to modify the existing network to allow resource reservation to provide performance guarantees. An example of this approach is the resource ReSerVation Protocol (RSVP) [Zhang et al. 1993, Braun 1998]. These have not yet been deployed on a large scale due to the high expense involved. Hence, the transmission must be done within the constraints of the availability of the bandwidth. If the transmission rate is less than the available bandwidth, it results in sub-optimal video quality and if the transmission rate is more than the available bandwidth, then routers are not able to cope with the load leading to congestion, resulting in packet drops and hence degradation of quality. Therefore, the available bandwidth and the transmission rate should be well matched. This can be done by first estimating the availability of the bandwidth, and then matching the transmission rate according to the available bandwidth [Bolot and Turletti 1998].
In order to estimate the availability of the bandwidth there are basically two types of methods:

- **Probe-based method.** In this method the sender carries out probing experiments in order to estimate the bandwidth. For example, the sender may keep on increasing the transmission rate till the packet loss rate increases beyond the set threshold. The advantage of this technique lies in its simplicity. There are a number of methods like Variable Packet Size (VPS) probing, Packet Pair/Train Dispersion (PPTD) probing, Self Loading Periodic Streams (SLoPS), Trains of Packet Pairs (TOPP) for this task [Wu et al. 2001, Prasad et al. 2003]
• **Model-based methods.** In this method the aim is to make the transmission as fair to the existing TCP flow as possible. Thus it is also called as the “TCP friendly” method. In this, the average TCP flow is modelled and then the video transmission is controlled in accordance with that. Hence, the video streaming and the TCP flow match on the macroscopic level [Wu et al. 2001].

Once the availability of the bandwidth has been estimated, the sender can adapt the transmission to the available bitrate by adapting the video bitrate. The goal of a rate control algorithm is to control the bitrate to match the network requirements while minimising the reduction or variation in visual quality of decoded video sequences [Ghanbari 1989, Paul et al. 1998, Johanson 1999, Lee and Kim 2000, Viéron et al. 2002, Setton et al. 2003, Qian and Hwang 2003]. The rate control can be performed either at the sender end or the receiver end and a number of techniques have been proposed to do this.

In a sender-controlled adaptation, receivers send feedback to the sender regarding their respective bandwidth availability, and based on the feedback, the sender explicitly adapts the transmission rate [Bolot and Turletti 1998, Braun 1998, Smith et al. 1998]. In general, since the sender controlled adaptation is carried out using the average of all receivers’ network state, it is difficult to cater to the demands of an individual receiver [Na and Ahn 1999]. The INRIA Videoconferencing System (IVS) is a multicasting system running over the RTP/RTCP stack and is an example of a tool that employs sender controlled adaptation techniques. In this system the sender determines the network state by averaging the packet losses suffered by the receivers and then adjusts the transmission rate by linearly increasing or decreasing the rate [Turletti 1994]. IVS suffers from the problems of excessive packet drops, unfair sharing of the network with other traffic sources because of the inaccurate evaluation of the network state, static rate control algorithm, delayed feedback. Overall, it cannot effectively handle short term congestions [Na and Ahn 1999]. An improvement in the architecture has been suggested where the condition of the network is judged not only by the packet losses but also by the variations in the round trip times [Na and Ahn 1999].

In the receiver initiated adaptation, the sender sends the video in a number of channels or layers and the receiver regulates the receiving rate by adding or dropping layers [Wu et al. 2001].
A number of rate control algorithms have been developed for MPEG. Most of these algorithms consist of two parts. The first part smoothens the variations arising due to the MPEG coding in the form of GOPs by using a buffer at the output of the encoder. The second part adjusts the long-term variations in the quality of video by adjusting a variety of parameters [Bolot and Turlleti 1994, Zhao 2001, Kim and Altunbasak 2001]:

- Frame rate
- Spatial resolution
- Quantisation
- Movement detection threshold
- Group of Pictures (GOP) structure
- Refresh rate

The control of frame rate and resolution is an easy and effective way of regulating the transmission rate. In a quality driven adaptation algorithm called DAVE [Nepal and Srinivasan 2003], the quality of video is identified and then the frame rate and resolution adjusted so as to deliver the video within the available network bandwidth. In the prototype model, the algorithm tries to estimate the spatial and temporal resolution of the video sequence and then adjusts the codec (MPEG 7) to transmit the perceptually best video within the available bandwidth. In the Video Transport Protocol (VTP) for MPEG-4 [Balk et al. 2004], the sender first estimates the available bandwidth and then transmits the video sequence. If the receivers are able to accept the sequence at the rate transmitted, the sender then increases the rate to probe the network for unused bandwidth. If an increase leads to packet drops then it lowers its transmission rate. By continuously probing the network condition the sender adjusts its transmission rate. The transmission rate is adjusted by adjusting the frame size keeping the frame rate and the Group of Video Object (GOV) planes constant.

In MPEG-4, each macroblock is processed by first applying the discrete cosine transformation (DCT) and then quantising it. Quantization can be carried out between the range 1 to 31 where a higher number indicates higher compression, lower quality and lesser bandwidth requirements. It is thus possible to adjust the transmission rate by adjusting the quantisation value. In an advancement to this technique, it has been proposed [Zhang Q. et al. 2001] that the background and foreground be first separated and then transmitted on different VOPs. Depending upon the network condition, the receivers send feedback to the sender, which can adjust the transmission rate by adjusting the quantisation parameters (QPs) independently.
for the different VOPs. By this technique higher compression can be applied to the background and lower to the foreground.

MPEG codecs identify the foreground and background based on the degree of change in the macroblock. The macroblocks experiencing significant changes are considered part of foreground and transmitted while the background macroblocks are transmitted less frequently. Hence, by adjusting the threshold of the change detection the rate can be adjusted [Bolot and Turletti 1998].

Refresh rate is the fraction of frames that are accepted by codec from camera for encoding and transmission. By controlling the refresh rate, the frame rate can be controlled thereby controlling the bandwidth [Bolot and Turletti 1994].

The adaptation needs to be based on the type of video sequence. For example for head-and-shoulder video, which has low motion content, temporal scaling (reduced frame rate) is preferred over spatial scaling [Masry et al. 2001]. On the other hand, in case of high motion sequence, high frame rate should be maintained while reducing the spatial and SNR scaling [Kim and Altunbasak 2001]. When the perception of movement is important then the quality should not be adjusted by changing the quantisation and movement detection threshold [Bolot and Turletti 1998]. A systematic study of the effect of frame rate and resolution on the perceptual quality of the VQEG test sequences have shown that the viewers prefer low frame rate for poor quality, low motion or high detail sequences and high frame rate for high panning or high movement sequences. In general a fixed 15 fps seems to be a reasonable compromise for all types of sequences [Yadavalli et al. 2003].

2.3.3 Layered Video Transmission

One of the technical challenges is to provide a service to users connected by links having different access bandwidths and terminal device processing capabilities. While multicasting to such heterogeneous networks it is important to ensure that all the users receive data at a rate appropriate to their own capabilities regardless of the capacities of other users.

Over the years a number of approaches have been researched in order to solve the challenge of satisfying users having diverse capabilities. These approaches can be divided into three types [Kim and Ammar 2001]:

- **The replicated stream approach**, in which the sender multicasts several streams with identical contents but at different rates and the receiver subscribes to the most suitable stream.
The cumulative layering approach, in which the video is partitioned into base and enhancement layers. The base layer contains the basic information of the image that can be accepted by all terminals, and the enhancement layers provide further refinement to the basic quality of the image. At the receiving end, these layers can be combined, depending on the bandwidth limitation and the processing capabilities of the receivers to develop the program. The quality of the presented image depends upon the number of layers combined. Since the Internet does not support layering it has not been practically implemented.

The non-cumulative layering approach, in which the video is encoded in two or more independent layers and the receiver can join any number of layers without necessarily joining the base layer.

A number of protocols have been proposed in order to send video in a layered format. These protocols can be roughly divided into two main categories: receiver-initiated and sender-initiated. In the receiver-based adaptation technique, the sender transmits the video at a constant rate [Ghanbari 1989, Li et al. 1999a, Lavington et al. 2000, Legout and Biersack 2000b]. The task of the receiver is to dynamically judge the number of layers it can receive and combine the layers (Figure 2.9).
Chapter 2

The classical receiver-initiated solution that has been proposed is the Receiver-driven Layered Multicast (RLM) approach [McCanne 1996, McCanne et al. 1996, McCanne et al. 1997]. RLM operates within the Internet protocol architecture and multicasts as best effort, multi-point packet delivery. To drive the adaptation, a receiver must determine if its current level of subscription is optimum. By definition, the subscription is high if it causes congestion and is easy to detect because congestion is expressed explicitly in the data stream through lost packets and degraded quality. On the other hand, in order to move from low level to higher level, the receivers carry out "join experiments". RLM suffers from a number of problems like sluggish dropping of layers, long time requirement for adjusting to higher layer, unfair sharing of bandwidth, conservative utilisation of bandwidth resulting in low link utilisation, instable quality, etc. [Legout and Biersack 2000b, Li et al. 1999b, Gopanakrishnan et al. 1999].

A modification of RLM is the Layered Video Multicast with Retransmission (LVMR) protocol [Li et al. 1997, Li et al. 1998, Paul et al. 1998, Li et al. 1999a, Li et al. 1999b]. In this approach, video quality has been improved by intelligently retransmitting lost packets and by the receivers dynamically adjusting the video reception rate using a hierarchical approach. In the hierarchical approach used in LVMR, there are multiple subnets within a multicast group and each subnet has a subnet agent. The subnets are further divided into domains, which can be either physical regions, a geographical area, or a logical region like an intranet etc. There is an Intermediate Agent (IA) within each domain, which is responsible for gathering information regarding the status of the domain. The information provided by the subnet agents is compiled by the IA, and passed down to the receivers. The information flow is in two directions in the whole hierarchy. The subnets and intermediate agents can be logically located, or can physically be on the same machine as the receiver.

In the sender-initiated approach the sender multicasts video streams whose quality is adjusted based on the feedback information from the receivers. One of the protocols which are based on the sender-controlled transmission is Source Adaptive Multi-Layered Multicast (SAMM) [Vickers et al. 1998, Vickers et al. 1999, Vickers et al. 2001]. It is an end-to-end source-adaptive multi-layered multicast. SAMM sends the video over a number of layers but the selection of how many video layers to send is dependent on the feedback from the receivers, but feedback from many receivers results in feedback implosion. To prevent feedback implosion the concept of feedback merger, in which the number of requests is reduced by is used.
Since, both the schemes have their respective advantages and disadvantages there are certain hybrid approaches that combine the good points of both. One of the problems with receiver controlled layering is that the adaptation is done only at the receiver’s side and the sender's side has a static rate allocation. As a result, there is a significant mismatch between the fixed transmission rates and the heterogeneous and dynamic rate requirements from the receivers. This mismatch can be minimized by employing dynamic layer rate allocation at the sender’s side. Hybrid protocols like Hybrid Adaptation Protocol for Layered Multicast (HALM) [Liu et al. 2002] combine the strengths of sender-initiated and receiver initiated approaches. HALM uses pure end-to-end control in which a metric, called Fairness Index, is provided at the sender end which reflects the degree of satisfaction of the receiver. This helps in deciding the optimal number of layers and the rate at which these layers are to be sent. [Salamatian et al. 2001, Liu et al. 2002].

In another hybrid approach, Layered Multicast Control Protocol (LMCP), the receivers not only select the number of layers they can accept but also send feedback to the sender. Based on the feedbacks, the sender then adjusts the transmission rate of each layer [Smith et al. 1998, Smith et al. 1999].

Similarly SIM (Selective participation, Intra-group transmission adjustment and Menu adaptation) combines the features of RLM (selective participation by joining the number of layers based on ability) and SAMM (Intra-group transmission adjustment of rates of layers) with menu-adaptation mechanism (selection of appropriate group transmission rates that improve overall efficiency of the session) [Gorinsky et al. 2001].

2.3.4 Basis for Layering

The layering can be done inside the codec or as a post-processing filter at the system level. In the former case the layering is done by the codec itself and is dependent upon the type of codec. While some of the codecs support layering, others do not. Layered video transmission is not supported by the H.261 codec, but it can be achieved by either modifying the codec or by adding a filter after codec. While H.263 does not support layering, the extended version of (H.263+ or H.263 Version 2) supports up to 15 layers. MPEG-2 supports layered representation but does not work efficiently at low bit rates because it relies on intra-frame updates to re-synchronize the decoder in the presence of errors or packet loss. In the latter case, a simple technique is used in which a filter parses through the output stream and reads the header. Depending upon the type of layering required the packets are directed to the appropriate multicast group. At the decoder, a multiplexer sequences video data from the different multicast groups and sends a stream to a decoder where it is decoded.
The number of layers into which the transmitted video can be partitioned is either static or dynamic. In the receiver-initiated schemes the number of layers is static while in the sender-initiated schemes it is dynamic. For example, in the RLM [McCanne 1996, McCanne et al. 1996, McCanne et al. 1997], the authors considered a static 4 layer system; on the other hand, in HALM (Hybrid Adaptation Layered Multicast) [Liu et al. 2002] the sender dynamically optimises the number of layers. By simulation, it was found that while the performance of HALM increases with increase in the number of layers, there is no significant increase after 5 layers [Liu et al. 2002]. In LMCP the transmission was done in ten layers and it was found while the performance improves with an increase in the number of layers, beyond 6 layers there was no significant improvement in terms of bandwidth utilisation [Smith et al. 1999]. SIM was tested in 5 layers because the number was thought to be appropriate [Gorinsky et al. 2001]. In the ThinStreams protocol, the layers (thick streams) are further divided into layers (thin streams) so that there are a number of layers. In the simulation carried out using 25 thin streams it was found that such a large number of streams resulted in considerable increase in overhead, increase in convergence time but helped to decrease quality oscillation [Wu et al. 1997].

The suitable way for partitioning the video stream should be judged based on the satisfaction of the users and the requirements of the network. To judge user satisfaction, tests need to be carried out on the perceived quality of video [Thakur et al. 2001]. A layering technique would be considered as network “friendly” if it satisfies the following criteria:

1. The coding is simple enough to allow minimal real time processing by the sender and decoding is of low-complexity and requires low-memory so that relatively unsophisticated devices can be used.

2. The scheme is robust enough to handle wide differences in the heterogeneous environment and also handle unpredictable and dynamic network bandwidth changes in the network conditions.

3. It is resilient enough to handle some packet losses during the transmission.

4. It is able to handle both unicast and multicast applications.

5. As far as possible, each layer is mutually independent. If that is not possible then the higher layer is dependent only on the lower layers.
Taking into consideration these points a good layering technique should consider the following four aspect of layering:

- **Coding Efficiency** can be defined as the ratio of the number of bits used by the non-layered codec to the number of bits used by the layered codec to achieve the same quality of picture. In some cases, the efficiency may fall drastically with the number of layers [Handley and Jacobson 1998].

- **Coding Complexity** is the additional calculations that must be performed by the codec in order to partition the video into layers. With the increase in complexity the coder and decoder would require additional time, resulting in poor video quality [Handley and Jacobson 1998].

- **Ability to Handle Heterogeneity** is the upper and lower limit of bandwidth that can be satisfied by the layered stream. By increasing in the number of layers, it is possible to satisfy greater heterogeneity, but the range of bandwidth covered does not increase linearly with the number of layers [Deering 1991].

- **Perceptual Quality of Video** normally increases with more layers, but the quality does not increase linearly. It is even possible that the quality may fall with an increase in the number of layers because of increased complexity and inefficiency. This is especially true for layering based on SNR [Handley and Jacobson 1998, Handley et al. 1999].

2.3.3 Techniques for the Division of Video into Layers

Layering of video data for transmission can be achieved by any of the following four layering techniques currently available. The performance of these layering techniques depends upon the codec used. In the coding community, scalable coding is frequently used to refer to layered coding. The scalability can be achieved by scaling the frame speed (temporal scalability), frame size (spatial scalability) and frame quality (quality or SNR scalability) or a combination of these.

**Temporal Layering**

In this type of layering the Intra (I), Predictive (P) and Bi-directional (B) video frames are sent on different layers. For example, in developing Layered Video Multicast with Retransmission (LVMR) [Zhang et al. 2001] the simulation tests were carried out for transmitting video in four layers using MPEG-2, where I frames made the base layer, P frames made the first enhancement layer, and B1, B3, B5 and B7 made the second enhancement layer whilst B2, B4, B6 and B8 made the third enhancement layer. I-frames can be independently
decoded, while P-frames require I-frames, and B-frames generally require both I and P-frames for decoding. During congestion, the order of preference for the dropping frames should in general be first B, then P and finally the I-frames this is because when the B-frame is dropped it influences the quality of one frame only, but when one P frame is dropped it influences the current frame and all the consequent B frames, but when I frame is dropped it influences all the group of pictures (GoP) and also the B frames of the previous GoPs. [Zhang et al. 2001]. It has been shown that the amount of impairments increases at twice the rate when I frames are dropped compared to when the P frames are dropped [Zhang et al. 2001]. This is not always true since experiments have shown that dropping some particular B-frames may result in poorer performance than the dropping of some P-frames. Hence, there is a preferential order of frame drops that should be determined for achieving the best result [Paul et al. 1998].

Temporal layering is easiest to implement and introduces almost no overhead because it can be done at the post-codec level by reading the header field to identify the frame type. The problem with this layering technique is that a reduction in layer numbers leads to severe degradation in the quality. If there is a loss of packets in the lower layers then some packets in the upper layer become unusable because the pictures in the higher layer are based on those from the lower layer [Kimura et al. 1999].

**Spatial Layering**

Spatial layering can be achieved by various ways including scaling the DCT coefficients, pyramid coding and spatial sub-band coding. Scaling of DCT coefficients can be done by various methods like [Amir et al. 1996, Kimura et al. 1999, Zhang and Xu 1999]:

- **Layered quantization**, in which an 8x8 block of each image is transformed into the frequency domain and the DCT coefficient magnitudes are partitioned into different layers with more significant ones sent to lower layers.

- **Spectral separation**, in which the video is split between a number of layers, based on the spatial frequencies. Since the lower frequencies are better seen by humans as compared to the higher ones, hence they are sent in lower layer while the higher frequencies are sent in higher layers.

- **Spatial scaling**, in which the spatial resolution of the image is increased from the lower to higher layers.
In pyramid coding the encoder first down-samples the image, compresses it using the required encoder and transmits it in the base layer. When the image is decompressed and up-sampled a much coarser copy of the original is obtained. The difference between the original and the up-sampled image is sent in the enhancement layer [Kuhmunch et al. 2001].

A problem with the spatial layering is the tendency of higher layers to depend on the lower layers; hence, if there is packet loss in the lower layers, then the distortion tends to get propagated to the higher layers [Zhang and Xu 1999].

**Data Partitioning (DP) layering**

In the Data Partitioning (DP) layering technique, the layering is performed by allocating specific bytes in the bit stream (i.e. motion vector information and Discrete Cosine Transform (DCT) coefficients) into different layers. For example MPEG-2 has Priority Break Points (PBPs) in each slice header that specifies the grouping of DCT coefficients into the layers. The layering method using these PBPs, the vectors and coefficients with higher importance can be directed to the base layer and the rest to the enhancement layers. It has been seen for MPEG-2 that the quality of picture increases through 4 layers, beyond this, there is not much increase. It has also been seen that there is a sequence of frame drop which gives the best performance. For example, if all B frames are dropped prior to dropping I and P frames then the quality loss is disproportionately high [Kimura et al. 1999].

After Temporal, DP layering is the easiest to implement because it requires only a post-processor filter. The efficiency of this method is also high since only headers (namely, sequence, group of pictures, picture and slice headers) are included in all layers [Kimura et al. 1999].

The major disadvantage of the DP layering is that even a small number of packet drops in the base layer degrades the quality significantly. It has been seen that the quality falls almost linearly with the number of packets lost in the enhancement layers and the enhancement layers are dependent upon the base layer. Therefore, the base layer should be protected at the expense of the enhancement layers [Kimura et al. 1999].

**Signal to Noise Ratio (SNR) Layering**

SNR layering is done at the codec level by encoding the video using a quantiser scale in order to generate a base layer. The enhancement layer is later generated by encoding the difference between the original video and the base layer using a quantiser scale. For more than two layers, the same process is applied recursively using additional quantiser scale parameters. For SNR layering, the quality degrades linearly with the rate of packet loss in the enhancement
layer and drops rapidly when the data loss occurs in the base layer [Kimura et al 1999, Lavington et al. 2000, Amon et al. 2002].

In this type of layering the efficiency is rather low because the DCT coefficients are divided among the layers, and information for every DCT coefficient is included in all the layers. The efficiency falls rapidly with an increase in the number of layers, so it is not practical to have more than 2 layers [Kimura et al 1999].

**Fine-Granular Scalability**

Studies have been carried out to exploit the developments in the MPEG-4 fine granular scalability (FGS) for Internet applications. In fine granular scalability the receiver can start decoding and displaying even after receiving small amount of data, and as more data is received the quality is progressively improved. In this system the base layer is sent so as to provide the basic coarse quality. On top of this base layer an enhancement layer based on FGS is sent. Due to the fine-granularity of the enhancement layer, the sender can control the adaptation of the bit-rate depending upon the network condition. This is done by encoding and transmitting a part or whole of the enhancement layer [Radha et al. 1999, Wu et al. 2001].

**Combination layering**

All the different types of layered coding schemes described above perform well according to their restricted premise, but the quality of video can be improved further by combing or exploiting the feature of multiple layering techniques. For example, this can be achieved by a trade-off between temporal and spatial resolution. Thus in the scenes involving high motion the stress is on the increased number of frames at the expense of resolution, hence more coarse frames should be sent on the lower layers and the resolution refinement should be sent on the higher layers. On the other hand in the case of scenes involving little motion the few frames having high resolution should be sent in lower layers and more frames should be sent in higher layers.

2.3.c **Summary of the Literature Review for this Section**

From the literature survey carried out it may be concluded that:

1. The transmission of video is sensitive to packet drops and delays, hence, they need to be minimised by suitable rate shaping algorithms. This rate shaping can be done by controlling the quantisation, frame rate, resolution and/or GOP structure.
2. Layering techniques for video transmission are a promising way to satisfy receivers in heterogeneous environments and a good layering technique is required for efficient partitioning.

The literature survey has shown the need for rate shaping and efficient compression of video sequences before transmission. The next section discusses compression techniques based on the separation of background and foreground.

2.4 VIDEO COMPRESSION TECHNIQUES BASED ON THE SEPARATION OF BACKGROUND AND FOREGROUND

The separation of background and foreground followed by the transmission of the foreground at high quality and the background at relatively low quality is a common compression technique adopted by most of the video codecs, to reduce the data that has to be transmitted over the internet. While a number of papers have been written on the separation of foreground and background, none of them have actually defined the two words. An extensive search on the internet also did not clearly distinguish between the two.

According to the Merriam-Webster Online Dictionary:

**Foreground:**

1. The part of a scene or representation that is nearest to and in front of the spectator

2. A position of prominence

**Background**

1.a. The scenery or ground behind something

1.b The part of a painting representing what lies behind objects in the foreground

2. An inconspicuous position

According to the American Heritage Dictionary:

**Foreground:**

1. The part of a scene or picture that is nearest to and in front of the viewer.

**Background:**

1. The ground or scenery located behind something.

2.a. The part of a pictorial representation that appears to be in the distance and that provides relief for the principal objects in the foreground.
2. b. The general scene or surface against which designs, patterns, or figures are represented or viewed.

3. A position or area of relative inconspicuousness or unimportance.

Based on these definitions it is difficult to tell which part of a picture is a foreground and which is a background. Hence, various researchers have tried to identify them based either on the temporal or the spatial differences between the two.

2.4.a Identification of Foreground and Background Based on Temporal Differences

The identification of “background” and “foreground” based on the temporal differences has been adapted by a number of codecs. For example, in the MPEG codecs, one way to identify and remove background (often considered as the redundant data) is by the concept of conditional replenishment. In a frame, which is made up of blocks of 8 x 8 or 16 x 16 pixels, only those blocks that change substantially in successive frames are encoded and sent. One of the ways by which the conditional replenishment works is by using the “block ageing” algorithm [Kimura et al. 1999]. According to this, when motion is detected in a block, the block is sent as a low quality block, if there is no motion in the block, the block is kept under an aging process. If the block remains unchanged for “threshold” time then the block reaches the Idle State. Some of the idle blocks in each frame are sent to Background State (BG). The background blocks are refreshed less frequently using high quality blocks. During the ageing process, if a block undergoes changes, then the block goes back from Idle / Background State to the initial Motion State [Puri and Chen 2000]. Conditional replenishment is significantly simpler than other video compression methods in terms of computational complexity, while still able to achieve an acceptable compression ratio for certain video applications. Experiments with desktop video without motion compensation reveal that on average, less than 3% of the pixels need to be replenished in most head-and-shoulder scenes with CIF/Q-NTSC frame sizes.

The problem with the conditional replenishment scheme is in defining what constitutes a “change or motion”, and for this too, a number of schemes have been proposed. One of the perception-based change detection methods is based on the contrast sensitivity of human visual system; the authors suggest that only the pixel having intensity differences relative to the stimulus background which exceed the visibility threshold, should be encoded. Hence, different regions in the frame are defined as Background and Foreground, based on the degree of motion. This change or motion detection is the weakness of the conditional replenishment scheme because improper change detection leads to blockiness [Chiu and Berger 1996].
2.4. b Identification of Foreground and Background Based on Spatial Differences

Researchers working in the area of pattern recognition have tried to distinguish between the foreground and the background, based on the spatial differences like, gray scale luminosity, colour scheme, etc. This can often lead to misleading results. For example, in Figure 2.10, an illustration by M. C. Escher, the gray scale luminosity will not be able to distinguish between the two because the roles of black and white have reversed. Similarly, Salvador Dali in his painting Slave Market / Voltaire has varied the colour scheme so that the bust of Voltaire and the heads of slave girls are interchangeable (Figure 2.11). These examples illustrate the difficulty in identifying foreground and background based only on spatial differences.

Psychologists have also studied the identification of “background” and “foreground” under the topic “figure-ground phenomenon”, based on spatial differences. According to them, the figure-ground relationship is an important element of the way the understanding of the scene is organised in the human brain. According to the New Fontana Dictionary of Modern Thought [2000] the figure-ground phenomenon is the characteristic organization of perception into a figure that 'stands out' against an undifferentiated background, e.g. a printed word against a background page. What is figural at any one moment depends on patterns of sensory stimulation and on the momentary interests of the perceiver.

Edgar Rubin, a Danish psychologist, was the first to systematically investigate the figure-ground phenomenon. The phenomenon illustrates that in perceiving an image, some objects take a prominent role (the figures), while others recede into the background (the ground). This is illustrated by the classical Rubin Face/Vase Illusion (Figure 2.12), which is named after him. It is important to know that eyes can focus on only one “interpretation” at a time. It is not possible to observe both the figure and ground at the same time. It has been proposed that the distinction between “figure” and “ground” is due to the fact that perception is relative rather than absolute. Thus, eyes and brain register changes and relationships in the world, rather than the absolute quantities. Humans perceive a “foreground” as “bright” / “distant”/ “colourful”, as compared to something else “background”. Hence, foreground and background are not absolute but depend upon the conditions of perception.
FIGURE 2.10. Escher’s wood engraving.

FIGURE 2.11. Salvador Dali’s Slave Market / Voltaire

FIGURE 2.12. The Rubin Face/Vase illusion
2.4.c Summary of the Literature Review for this Section

From this brief literature survey the following points stand out:

1. It is possible to save on bandwidth by the separate transmission of background and foreground, wherein the background is transmitted in poor quality or infrequently.

2. There is a need to find better techniques for the identification of the two – background and foreground – portions in a scene.

3. Understanding of the human visual system can provide a basis for the identification of the two components.

4. There is a need to select or develop appropriate algorithm(s) for separating the two portions.

Since the understanding of the human visual system can provide a basis for the identification and separation of the foreground and background, the basics of the human visual system shall be discussed in the next section.

2.5 BASICS OF THE HUMAN VISUAL SYSTEM

Visual perception is one of the five senses. It consists of the ability to detect light and to interpret (see) it as the perception or vision. This perception, or vision, is brought about by a specific sensory system, the visual system. The visual system consists of two parts:

- The physical or the sensory part (the hardware).
- The psychological or the cognition part (the software).

2.5.a The Sensory Part of Human Visual System

The eye is the light-sensitive organ and is the first component of the visual system (Figure 2.13). The eye’s retina, which covers 72% of the inner wall, consists of photoreceptor cells – rods and cones. It performs the first stages of visual perception processing – converting the light into neural signals. These signals are further processed by other neurons of the retina and then transmitted through the optic nerve, made up of the signal carrying cells, ganglions, to the visual cortex of the brain [Winkler 1999].
The rods respond to dim light and mediate lower-resolution, black-and-white, night vision. They get easily saturated in bright light. On the other hand, the cones function in brighter lighting because they are not sensitive enough to work at very low light levels and mediate high-resolution vision and colour vision. Rod and cone signals are intermixed and combine to form the complete picture [Winkler 1999].

At the centre of the eye is the fovea, which is most sensitive to light and is responsible for the sharp central vision. This sharp central vision is necessary when reading, watching video, or any other activity where visual detail is of primary importance. At the centre of the fovea there is a pit with a diameter of about 0.2 mm. It has a high concentration of cone cells and virtually no rods. As a result, the fovea is responsible for highly detailed vision but is not sensitive to dim lights. The density of receptor cells decreases with increase in the distance from fovea, and additionally the rod cells become more predominant at the periphery. Due to the low density of receptor cells and high proportion of the rods as compared to the cones, the peripheral vision is weak in humans, especially at distinguishing colour and shape; but it is good at detecting motion and is relatively strong at night or in the dark, when the lack of colour cues and lighting makes cone cells far less useful [Winkler 1999].
Chapter 2 Literature Review

2.5. b Visual Attention

An image that is produced by the excitation of the retinal receptors (the cones and rods) is further processed by the neuronal system and various parts of the brain working in parallel to form a representation of the external environment in the brain. To perform its task, visual perception takes into account not only patterns of illumination on the retina, but also input from other senses and past experiences [Cater et al. 2002, Cater et al. 2003b]. For example, Figure 2.14 shows the photographs of different moths in camouflage, yet human eyes can identify them as “moths”. In this identification of moths against a background of leaf, flower or tree bark, one needs a prior exposure to general properties of moths. From past experiences, one expects the moths to have a certain shape, colour, etc. and then the eyes “search” for the shape.

An eye normally views a wide area, containing a number of objects. Such a complex scene is difficult to process and remember. Hence, the human eye only focuses on the objects that fall on the fovea, and when the detailed information of a large scene is required, the eyes redirect the gaze so that the relevant sections of the scene fall sequentially on the fovea. This redirecting of the eye, on the objects of interest is called eye saccade. At each jump a very small portion of the whole scene falls on the fovea to produce a detailed image of the spot. The brain then reassembles these glimpses into a coherent, but inevitably imperfect visual perception of the environment. In the process, a number of details get lost [Cater et al. 2002, Cater et al. 2003b].

The use of saccades for covering a scene was first studied in detail by the Russian psychologist Yarbus. In a classical study he asked the subjects to view Repin’s painting “An Unexpected Visitor” to demonstrate the task dependence of eye movement pattern. Using a crude method for eye tracking, optical stalks attached to the sclera via suction, he demonstrated that the subjects’ eye movement records varied dramatically with instruction (Figure 2.15).

FIGURE 2.14. Photographs of different types of moths in camouflage, showing that in spite of all the variations in size, shape and setting, human eyes can still recognise the moth.
1. Free viewing.  
2. Judge their ages.  
3. Guess what they had been doing before the unexpected visitor's arrival.  
4. Remember the clothes worn by the people.  
5. Remember the position of the people and objects in the room.  
6. Estimate how long the unexpected visitor had been away from the family.

**FIGURE 2.15.** Effects of task on eye movements. Repin’s picture was examined by subjects provided with different instructions [Cater et al. 2002, Chalmers et al. 2003].
FIGURE 2.16. Demonstration of the bottom-up processing by the standing-out of fruits from the rest of the painting in Paul Cézanne’s “Apples, Peaches, Pears and Grapes”.

He showed that once the appropriate object has been identified by the initial eye saccade, the eye performs a smooth movement to keep the object of interest in the foveal vision. The object that is followed by the eye thus has sharp image and the peripheral objects are viewed as smeared and unclear [Cater et al. 2002]. Thus, while performing a task, attention is focussed only on a small region at a particular time [Cater et al. 2003b]. This visual attention is coordinated by conscious and unconscious processes in the brain, which allow the subject to find and focus on relevant information quickly and efficiently [Cater, 2003a]. Thus, visual attention is the process by which humans select a small portion of a scene for extracting detailed information and use it for identification, understanding or other purposes. This allows preferential processing of the objects of interest over the unimportant objects.

2.5.4 Inattentional Blindness

The visual attention process is of two kinds: bottom-up and top-down. In the bottom-up processing the eye is naturally attracted to the most prominent, colourful, different or moving object. In this processing the attention is not dependent upon the observer’s knowledge of the stimulus; the stimulus itself provides the guidance. The examples of bottom-up processing includes, involuntary attention on the fruits in Paul Cézanne’s “Apples, Peaches, Pears and Grapes” (Figure 2.16), or the lips and eyes of another person since they are the most mobile and expressive elements of a face. This is an evolutionary development which helps in survival by finding food for meal or interpreting the expressions of the person in the front.
On the other hand, in top-down processing the eye are guided by the task at hand, this is done by voluntarily focussing attention on one or more objects that are relevant to the observer's goal when studying the scene. Examples of top-down processing include searching for a road sign or searching for a target in a video game, or searching for the camouflaged moths (Figure 2.14) [Cater et al. 2002, Connor et al. 2004, Wolfe 1999, Wolfe et al. 2003].

In top-down processing, when the subject is focussing attention on the task at hand, he/she may not pay attention to irrelevant, yet conspicuous objects in the scene. This is called Inattentional Blindness. Hence, Inattentional Blindness is the failure of the humans to see unattended items in a scene [Cater et al. 2002, Cater et al. 2003a].

This Inattentional Blindness has been studied in the context of rendering by Cater et al. [2002, 2003a]. They have found that when subjects are performing a task at hand, they only pay attention to small, but relevant sections of an image, and give far less attention to other irrelevant portions of the scene. Exploiting this behaviour, they demonstrated that the amount of computation required for generating a rendered image can be significantly reduced by only generating the objects of interest at high quality, and the rest at relatively low quality.

With this background knowledge, it is worth defining “foreground” and “background” in the context of e-learning video by studying which portions of a scene in an e-learning video would attract the maximum attention of the students. It is also worth studying what the reaction of the students would be when they are shown a video where the portions of interest are displayed at high resolution and the rest is displayed at relatively low resolution. Based on this it may be possible to reduce the bandwidth requirement by transmitting the important sections at a high resolution and the rest at lower resolution.

2.5.d Change Blindness

As a child, almost everyone has worked on the puzzle in which two similar drawings are displayed side by side. While they look similar, the two drawings have a specified number of differences between them and the puzzle is to find the differences in minimum time. While it looks simple, the puzzle requires considerable time to solve since it is quite difficult to notice the differences without consciously going through them from region to region. For a long time researchers have noted that people are quite poor at noticing such changes; but in recent years this subject has renewed considerable interest among researchers and has been researched as “change blindness”.

45
According to Simons and Rensink [2005], *Change blindness is the striking failure to see large changes that normally would be noticed easily.* It is caused when the change is separated by a disturbance. At times, even a large and repeatedly made change can go unnoticed for a long time. The disturbance between the scenes may be natural, like the eye movement (saccades) and eye blinks or induced artificially by image flicker; brief “splats” that do not actually cover the region of change. In the case of a movie sequence, the disturbance can be caused by a sudden or gradual change in scene. Once the difference has been spotted, the change becomes very obvious and eye catching [Rensink 2002, Simons and Rensink 2005, Turratto et al. 2002].

The research carried out during the past decade suggests that there are a number of important factors that can cause change blindness.

- **Attention.** Change blindness results when the change is unable to attract attention towards itself [Simons and Rensink 2005].

- **Expectation.** Change blindness is strong when the change is unexpected; this is because the attention is not directed towards the region when the change occurs, hence the observer is unable to catch it. [Simons and Rensink 2005].

- **Background or Foreground.** Changes to the central item are detected more readily than the changes to the background, even if the changes are of equal salience. This is because in general, the central item receives more attention as compared to the background [Turrattoa et al. 2002, Werner and Thies 2000]

- **Number of Changes.** Humans have the capability to catch only one change at a time; if more than one change occurs, then the change will be noticed one after the other and not at the same time. Since the attention can be focussed only at a small region at a time, it requires considerable expense of energy and time to catch more than one change.

- **Type of Change.** It has been shown that the changes in location are more difficult to detect than the changes in the identity, where identity is defined as the objects feature other than their position [Rich and Gillam 2000, Simons and Levin 1997].

- **Degree of Change.** It has been observed that if a change takes place in small steps, then it will go unnoticed until the net change has become large. On the other hand, if one large change is made, it is easier to catch it. For example, in an experiment the object was rotated at 1° per flicker, and it took about 30° change for fifty percent of the subjects to note the change; on the other hand, one change of 15° was easily noted. [Fernandez-Duque and Thornton 2003.]
Chapter 2

Object of Interest. It has been shown that the changes to the object of interest are easier to detect compared to the objects of marginal interest [Werner and Thies 2000].

Interest or Expertise of Subject. Individual interests and expertise play an important role in understanding the picture and focusing on the regions of interest. Prior knowledge or expertise of the situation depicted in the picture helps in focusing on the region of interest and selecting the meaningful unit out of the picture. It also helps in expecting the change at the particular region. Similarly, meaningful changes for an individual are easier to detect than non-meaningful ones. [Simons and Levin 1997, Werner and Thies 2000]

Number of Distractions. In the presence of distractions the identification of changes was difficult. The difficulty becomes more acute if the number of changes increases [Wright et al. 2002].

A number of models have been suggested to define human perceptual system and explain the existence of Change Blindness. The complexity of the task means that none of these models are able to explain all the observations convincingly. Some points do come out regarding the psychological aspect of observation [Bahrami 2003, Rensink 2000a, Rensink 2000b, Rensink 2002]:

- Under normal circumstances, a cursory look is given to complete visual field.
- The visual system can implicitly detect change, but in the absence of focused attention, the change does not reach awareness and consequently is not reported [Thornton and Fernandez-Duque 2000].
- From this visual field, a small region of interest is selected. This region of interest is then processed further for object recognition, transferred into memory, etc. Attention may change the visual representation so that things look different while attended.
- At any stage the conscious visual representation is composed of the "general" visual feel of the whole scene and the effects of attention on the object of interest.
- This visual representation has no memory. It exists solely in the present tense. When a visual stimulus is removed, its contents are lost to the visual system. Similarly, when attention is deployed away from some previously attended object or focus, no trace of the effects of attention remain in the visual representation.
During the making of a movie, different shots are taken at different times and settings. Hence, in spite of all the efforts, some continuity mistakes do creep in. In fact there are web pages dedicated to listing these mistakes in popular movies. One of the sites lists the best three continuity mistakes as [Rensink et al. 1997, Movie Mistakes 2006]:

- **Commando:** The yellow Porsche is totally wrecked on the left side, until Arnold drives it away, and it is fine.

- **Spider-Man:** In the scene where Mary Jane is being mugged by four men, Spider-Man throws two of the men into two windows behind Mary Jane. Then the camera goes back to Spider-Man beating up the other two guys. When the camera goes back to Mary Jane the two windows are intact.

- **Terminator 3: Rise of the Machines:** In the scene where John and Catherine are in the hangar at the runway, the Cessna's tail number is N3035C. When the plane is shown in the air, the number is N9373F. When they land, the tail number has changed back to N3035C.

These are quite glaring mistakes, but how many viewers have actually noticed them?

This knowledge of the change blindness in humans does raise some interesting possibilities of improving the quality of e-learning video transmission by exploiting the limitations of the human visual system. The question then arises whether viewers would be able to notice the discrepancies that may arise if the background and foreground are first separated, then transmitted separately in way that the foreground is continuously updated while the background is sent infrequently.

### 2.5.5 Summary of the Literature Review for this Section

From this brief literature survey the following points stand out:

1. The human eye only focuses on the objects that fall on the fovea, and when the detailed information of a large scene is required, the eyes redirect the gaze so that the relevant sections of the scene fall sequentially on the fovea.

2. The objects on which the eye will focus depend upon the task at hand.

3. Attention is paid to the object of interest, which is determined by the job being carried out.

4. Humans suffer from Inattentional Blindness, which is inability to see the unattended items in a scene.
5. Humans suffer from Change Blindness, which is the failure to see large changes that
normally would be noticed easily.

6. Inattentional Blindness and Change Blindness are more for objects of marginal interest
as compared to the objects of central interest.

The literature survey has shown that there is a need to improve the quality of synchronous e-
learning video. It has also been shown that the transmission quality may be improved by (a)
employing efficient compression techniques; and (b) adjusting the rate of transmission
according to the available bandwidth so as to minimise the packet drops and delays. While it is
reasonable to expect that the quality of transmitted video will improve as a result of the two
techniques; it is quite possible that there may be a loss of quality due to the generation of
artefacts due to these two processes. Hence, the next section discusses the generation of
artefacts.

2.6 DEGRADATION OF VIDEO DURING TRANSMISSION

During the compression and transmission of raw video the quality of video may fall. The
degree of alteration, compared to the original video, depends upon the quality of the codec,
operational parameters (quantisation value, frame rate, frame size, etc.) and the transmission
conditions. This fall in quality is due to the artefacts that are created at these two stages [Miras
2002].

2.6.a Encoding Artefacts.

Before being transmitted, the raw video goes through the codec where it is compressed to a
more manageable size. In the codec it passes through a number of stages like Discrete Cosine
Transformation and quantisation. At each of these stages some distortion and loss of fidelity
(like frame dropping) can take place. The maximum distortion takes place during the
quantisation stage. The main types of artefacts in a compressed video sequence are the

- **Blocking effect or tiling.** Blockiness is defined as a distortion of the image which is due
to the appearance of a basic block encoding structure. It is due to the independent
quantisation of blocks which results in discontinuities at the boundaries of adjacent
blocks. In this distortion, false horizontal and vertical edges are formed at the block
boundaries (Figure 2.17). Due to its pattern, this deformation is the most apparent visual
distortion. Blockiness can also happen in case of video sequences having high motions
Chapter 2 Literature Review

FIGURE 2.17. Blockiness or tiling in a coded image. (a) Coded image; (b) Edges of the blocks detected using Robert’s edge detector.

The image (a) has been enhanced using the “AutoBalance” feature of Microsoft Photo Editor. Image (b) has been enhanced to increase blockiness using “Autobalance” feature of Microsoft Photo Editor, edges detected using Robert’s edge detector, turned into negative and finally enhanced again using “AutoBalance” feature of Microsoft Photo Editor. Robert’s filter was used to highlight horizontal and vertical edges generated due to blockiness.

due to motion compensation failures. Blockiness becomes worse when a scene contains sections having high motion and also sections that are stationary [Masry and Hemani 2001].

- **Blurring.** Blurring is the reduction in the sharpness of edges and spatial detail due to the suppression of higher-frequency coefficients by a coarser quantisation (Figure 2.18). It is a global distortion over the entire image. Both blurring and blockiness lower the perceptual quality to the same level [Masry et al. 2001].

- **Temporal edge noise or mosquito effect.** This is defined as a form of edge business which manifests as a time-varying sharpness (shimmering) of edges of objects. This temporal artefact is the result of different coding of the same area of the image in subsequent frames.

- **Jagged motion.** Jagged motion is the result of poor motion estimation. When the residual error of motion prediction is large, then the quantisation error of the residual is also high.

- **Jerky motion.** Jerkiness results when the originally smooth and continuous motion is perceived as a series of discontinued images. This is often the case when the video is transmitted at low frame rates.
FIGURE 2.18. Blurring during coding. (a) Still from the original sequence. (b) Still from the coded sequence. (c) Edges detected using Sobel’s edge detector of the original frame. (d) Edges detected using Sobel’s edge detector of the transmitted frame.

The images have been enhanced using the “AutoBalance” feature of Microsoft Photo Editor. Images (c) and (d) have been turned into negative. Sobel filter was used to highlight blurring, the use of Robert’s filter would have highlighted the vertical and horizontal edges due to blockiness.
• Other artefacts are *colour bleeding* (smearing of colour between areas of strong chrominance difference), added *random noise, chrominance mismatch* (due to the use of the luminance motion vectors for chrominance motion compensation).

Some of the above effects are unique to block-based coding, while others are prevalent in other compression algorithms. For example, in wavelet codecs there are no block-related artefacts, as the transform is applied on the entire image, however such codecs suffer from increased blurring. Of these blurring is dependent upon the spatial resolution of the video and jerkiness is dependent upon the frame rate, all other artefacts are dependent upon the quality of the codec. The perceptual quality falls drastically due to these artefacts, but it is difficult to quantify the role of individual artefacts on the subjective evaluation. The role of these artefacts depends upon the type of sequence and the area of the frame. For example, an artefact in the foreground would significantly affect the perceptual quality compared to the artefact in the background [Masry and Hemami 2001, Masry et al. 2001].

### 2.6.b Transmission Artefacts

An important source of impairments is the transmission of the compressed video bitstream over the packet network. During streaming of the video, the video sequence is fragmented into a series of packets and then sent to the receivers. Depending upon the network conditions this can result in the following problems:

• **Packet Loss**: The loss of packets can lower the perceptual quality of the video and in critical conditions make the continuous playout impossible. Hence, most of the applications impose some limit on the packet loss to achieve acceptable visual quality. Despite the loss requirement, the Internet does not provide any loss guarantee. In particular, the packet loss ratio could be very high during network congestion, causing severe degradation of multimedia quality [Kuo 1998, Miras 2002, Rosenthal 2004].

A number of schemes have been propounded in order to improve the quality of video despite the packet loss. One of the schemes employs the use of Forward Error Correcting (FEC) code, in which some extra data packets are sent with the normal packet and whenever there is a packet loss at the receiving end, the receiver uses these extra packets to reconstruct the lost data packets. Even though this technique uses more bandwidth it is considered a good solution for the recovery of lost data packet. Examples of various error correcting codes are – Reed-Solomon-code, Tornado code and Carousel code [Bolot and Turletti 1996, Bolot et al. 1999, Byres et al. 1998, Tan and Zakhor 1999, Hong and Nosratinia 2002, Nguyen and Zakhor 2002]. An interesting idea that has been proposed is
the combination of FEC with layered coding [Karlsson 1997]. In this scheme the redundant data is sent into different layers and depending upon the requirement the receivers accept different number of layers.

The second way to correct packet drop is by retransmission of the lost packets. In this approach the receiver tells the sender which packets have been received or lost and the sender resends lost packets. Even if the network allows for re-transmission of lost packets, as is the case for wireless IP networks, the retransmitted packet must arrive before its playout time, where the playout time is the deadline by which the frames must be received / decoded and displayed. If the packet arrives too late for its playout time, the packet is useless and effectively lost. The advantage of the technique is that only the lost packets are sent hence the bandwidth is utilised more efficiently. The disadvantage is that it results in latency (round-trip-time (RTT)) and when the RTT is large its usefulness decreases. Its other limitation is the requirement of a back-channel which is not present in broadcast, multicast, or point-to-point without back-channel. There are a number of variations of this scheme, these include, retransmission only of the packets that can arrive in time and retransmission of important packets before unimportant packets [Kuo 1998, Miras 2002].

- **End-to-end delays:** Real-time video transmission is particularly sensitive to delays, hence every packet must arrive at the client before its playout time, with enough time to decode and display the packet. If a frame arrives after its playout time it is generally useless. In case subsequent frames are dependent upon the frame that has arrived late, then the delay gets propagated. The end-to-end delay in the arrival of the packets is the sum of the router processing and queuing delays, propagation delays, and end system processing delays. Since all of these delays are dynamic and unpredictable hence end-to-end may fluctuate from packet to packet [Kuo 1998, Miras 2002].

- **Jitter:** A variable delay on the IP networks is known as jitter. Figure 2.19 is the schematic representation of generation of a jitter. Jitter is mainly due to queuing and contention of packets at intermediate routers, but can also happen when packets take a different path to the destination [Miras 2002, Rosenthal 2004]. The perceptual quality of video falls almost as much due to jitters as due to lost packages. Interestingly, the perceptual quality falls very rapidly with low amount of jitter but does not fall proportionately higher with an increase in the level of jitter [Claypool 1999].
To some extent delay and jitter can be smoothened by adding a buffer at the decoder end. The addition of a buffer corresponds to the addition of an offset of about 5-15 s to the playout time of each packet. The addition of a buffer compensates for the delay / jitter and enables retransmission of lost packets. In this case the design of appropriate playout strategy is very important. While designing the strategy there has to be a trade-off between playout delay and late / lost packet arrival. Long playout delays lead to fewer late or lost packets. Normal streaming of stored video can tolerate long delays but real-time interactive videos cannot tolerate long delays [Kuo 1998, Miras 2002].

The impact of these artefacts depends on the nature of the video encoder and the level of redundancy present in the compressed bitstream (for example, intra-coded bitstreams are more resilient to loss). In MPEGs, due to the dependence of P frames on the previous frames, the loss of information can result in error propagation. This will continue until the decoder can re-synchronise. This is viewed as error blocks within the image (Figure 2.20); it bears no resemblance to the current scene and it usually contrasts greatly with adjacent blocks. Obviously, this has major impacts on perceived quality, and it is usually greater than the effects of coding artefacts [Kuo 1998, Miras 2002].
2.6.c Expected Artefacts in the Controlled Transmission of E-learning Videos

In this research it has been proposed that the quality of the transmitted e-learning video sequence shall be controlled by adjusting the frame rate and resolution. Among the encoding artefacts, blurring and jerky motion are expected to be the big problems, this is because all other encoding artefacts are inherent characteristics of the codec used and as such, independent of the two – frame rate and resolution. The generation of transmission artefacts is outside the scope of this work.

Figure 2.21 shows the effect of resolution on the blurring of the video sequence transmitted at the same resolution and at the 70% resolution. A comparison between Figures 2.21b and 2.21d shows that there is a slight deterioration in the quality of the transmitted sequence. This deterioration is particularly obvious for the fine portions of the image, like hair. The comparison between Figures 2.21b and 2.21f shows that there is a significant deterioration in the quality transmitted sequence when the sequence is sent at 70% resolution. This is manifested by a further blurring of the fine objects, like hair, and increase in the thickness of the border. Hence, the resolution has a strong influence on the clarity of the sequence.

Figures 2.22 and 2.23 show the effect of frame rate on the jerkiness of the motion. Figures 22a-22f are the six consecutive frames of a sequence transmitted at 25 fps and Figures 2.22g-2.22k are the difference between Figures 2.22b-2.22f and Figure 2.22a. Figures 2.22g-2.22k show a gradual change in the frames. As a result, when the Frames 2.22a-2.22f were played, the video sequence had a smooth movement. On the other hand, Figures 2.23a-2.23k are the six consecutive frames of a sequence transmitted at 6 fps. Figures 2.23g-2.23j are same and Figure 2.23k is quite different. Hence, when the Frames 2.23a-2.23f were played, the movement was quite jerky. Hence, the frame rate has a strong effect on the smoothness of the movement.

2.6.d Summary of the Literature Review for this Section

From this survey it is clear that:

1. The transmission of video over the Internet may result in some deterioration in the quality, primarily due to (a) the operating conditions and characteristics of the codec, and (b) the network conditions.

2. If the quality of transmitted e-learning videos is controlled by controlling the frame rate and resolution, then the deterioration in the quality is mainly due to blurring and jerky motion.
FIGURE 2.21. Effect of resolution on the blurring. (a) Still from the original sequence. (b) Edges detected using Robert’s edge detector of the frame from the original sequence. (c) Still from the sequence transmitted at 100% resolution. (d) Edges detected using Robert’s edge detector of the frame from the sequence transmitted at 100% resolution. (e) Still from the sequence transmitted at 70% resolution. (f) Edges detected using Robert’s edge detector of the frame from the sequence transmitted at 70% resolution.

The images have been enhanced using the “AutoBalance” feature of Microsoft Photo Editor. Images (b), (d) and (f) have been turned into negative.
FIGURE 2.22. Effect of frame rate (25fps) on jerkiness. (a) – (f) Six consecutive frames of a sequence. (g) – (k) Difference between the frames (b) – (f) and frame (a).

Images (g-k) have been turned into negative and enhanced using the “AutoBalance” feature of Microsoft Photo Editor.
FIGURE 2.23: Effect of frame rate (6 fps) on jerkiness. (a) - (f) Six consecutive frames of a sequence. (g) - (k) Difference between the frames (b) - (f) and frame (a).

Images (g-k) have been turned into negative and enhanced using the "AutoBalance" feature of Microsoft Photo Editor.
The scope of this research does not encompass the study of different codecs or the network conditions on the quality of video, hence the generation of only a few types of artefacts need to be considered. The measurement techniques used to quantify the quality of video should be able to account for these artefacts, hence the next section deals with the methods of characterising the quality of video.

2.7 PERCEPTUAL QUALITY MEASUREMENT

As discussed in the previous section, the transmission of video over the internet may result in some deterioration of quality. How the receiver will perceive the video will then depend upon a number of factors [Farias et al. 2003, Reibman et al. 2004]:

- The viewer’s eyesight, expectations, etc.
- The viewing conditions like the viewing distance and background lighting the choice of monitor and display conditions.
- The transmission condition like the actual video contents (the amount of motion and texture), encoding parameters (like overall bit-rate, compression algorithm, spatial and temporal resolution) and any subsequent modifications of the bit stream by either the server (intentional rate changes) or network (large delays, jitter, reduced throughput, or packet losses).
- Type of artefact, for example blurriness contributes the most to annoyance followed by noise and then blockiness.
- Decision made by the client in response to the received bits, including loss-concealment and client buffer strategies. Even in response to a bit stream received with no network impairments, the operating system may impact temporal smoothness.

The degree to which the quality falls has to be measured, which can either be done subjectively or objectively. Out of these the first two conditions can only be measured subjectively but the other three can be measured objectively. The sections below briefly discuss the various video quality measurement techniques that may be relevant to this study.
2.7. a Subjective Quality Measurements

Since it is the users who would ultimately decide the quality of video, the subjective tests are considered the "ultimate truth". Almost all the subjective tests are time-consuming and complex. This is because assessing the subjective perception of quality is a difficult task, and is influenced by a number of factors such as, frame rate, lighting, image size, movement, degree of synchronisation, and even clarity of audio [Winkler 1999, Lu et al. 2001]. The most widely-used methods for measuring the subjective quality of speech and video images have been standardised and recommended by the International Telecommunications Union (ITU). ITU-T Recommendation P.910 [1999] addresses subjective assessment of image quality over multimedia applications. It has recommended the following methodologies:

**Absolute Category Rating (ACR)** is a methodology in which the test sequences are presented one at a time and are rated independently on a category scale. At the end of each presentation the subjects are asked to evaluate the quality of the sequence shown. The subjects are asked to rate the quality on a five-level scale. [ITU-T Recommendation P.910 1999].

**Degradation Category Rating (DCR)** is a methodology in which the test sequences are presented in pairs. The first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. The subjects are asked to rate the quality on a five-level scale. The levels are [ITU-T Recommendation P.910 1999]:

1. Very annoying
2. Annoying
3. Slightly annoying
4. Perceptible but not annoying
5. Imperceptible

If higher discriminative power is required, a nine-level scale may be used.

The time pattern for the stimulus presentation can be illustrated by Figure 2.24. The voting time should be less than or equal to 10s, depending upon the voting mechanism used. The presentation time may be reduced or increased according to the content of the test material [ITU-T Recommendation P.910 1999].
Pair Comparison method (PC) implies that the test sequences are presented in pairs, consisting of the same sequence being presented first through one system under test and then through another system. The systems under test (A, B, C, etc.) are generally combined in all the possible $n(n-1)$ combinations AB, BA, CA, etc. Thus, all the pairs of sequences should be displayed in both the possible orders (e.g. AB, BA). After each pair a judgement is made on which element in a pair is preferred in the context of the test scenario. For the PC method, the number of replications generally need not be considered, because the method itself implies repeated presentation of the same conditions, although in different pairs [ITU-T Recommendation P.910 1999].

**2.7.b Objective Measurements**

Since the subjective tests are time consuming, hence a number of objective tests have been developed [Robertson and Fisher 1985.]. All these tests have their own strengths and weaknesses. The biggest weakness of all the tests has been that no single objective test has been able to correlate strongly with the subjective tests under all test conditions. In spite of this weakness, the objective tests are a very useful tool in giving an indication of the quality and have been found to be useful within their limitations.

The objective tests are basically of two types (Figure 2.25) [Cranley 2004]:

- **Tests that work in the absence of a reference sequence.** In this, the only test video sequence is used (Figure 2.25a). Its shortcoming is that in the absence of any reference clip, the quality measurement is subject to errors caused by picture content resembling the specific impairment parameters that are being detected.
• **Tests that work in the presence of a reference sequence.** In this method the reference video sequence and the test video sequence are fed to a computer algorithm that calculates the distortion between the two in every frame of the video sequence (Figure 2.25b).

### 2.7.c Commonly Used Objective Metrics

**Root-Mean-Square Error (RMSE)**

The Root-Mean-Square Error method is among the simplest objective test method. It calculates the pixel-to-pixel “difference” between two images. It can be applied to digital video by averaging the results for each frame. RMSE can be calculated as:

\[
RMSE = \sqrt{\frac{1}{M \times N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} [f'(m,n) - f(m,n)]^2}
\]

where
- \(M\) = total number of rows in the frame
- \(N\) = total number of columns in the frame
- \(m\) = row
- \(n\) = column
- \(f(m,n)\) = reference pixel at point \((m,n)\)
- \(f'(m,n)\) = test pixel at point \((m,n)\)
**Signal-to-Noise Ratio (SNR)**

SNR is the simple mathematical advancement of RMSE and can be calculated as:

\[
SNR = \frac{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f'(m, n)^2}{\sqrt{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} [f'(m, n) - f(m, n)]^2}}
\]  

(2.3)

**Peak-Signal-to-Noise Ratio (PSNR)**

The PSNR metric is calculated using only the luminance signal. An 8 bit image will contain pixel luminance values that vary from 0 (black) to 255 (white). It is yet another extension based on RMSE and can be calculated as follows:

\[
PSNR = 20 \times \log_{10} \left( \frac{255}{RMSE} \right)
\]  

(2.4)

The human vision is a very complex mechanism that has been researched for years. It cannot be correctly modelled using simple techniques, hence, the distortion calculated by an objective quality metric like RMSE might not match with the subjective perception of a human being. In other words, RMSE might calculate that the distortion between the original and impaired frame is high, but that might not affect the quality of the compressed video because the impairment occurred in a place that the human eye is less sensitive to.

The reasons for the failure of RMSE and all the other methods described earlier are that they do not take into account any of the following:

- Natural Motion
- Variant sensitivity of the human eye to contrast and spatial/temporal details
- Blockiness

**Spatial Information**

Spatial gradients, or edges, play important roles in image quality. There are a number of filters like Laplace, Prewitt, Kirsch, Roberts and Sobel that have been developed in order to detect the edges [Heath 1996, Wolf et al. 1997, ITU-T P.190 1999]. The values obtained by calculating the edges of the frames can be interpreted as an indication of the added or lost edges in the destination scene compared to the source scene. Added edges result from impairments such as tiling, error blocks, and noise. Lost edges can result from
impairments such as blurring. Either of these cases will result in a difference between successive frames and will increase the information that needs to be sent.

ITU-T P910 (09/99) has recommended the use of Sobel filter for the calculation of SI. In Sobel filter the edges are detected by linear convolving each video frame with the kernels [ITU-T Recommendation P.910 1999]:

**Vertical Convolution**

\[
H_v = \begin{bmatrix}
-1 & -2 & -1 \\
0 & 0 & 0 \\
1 & 2 & 1
\end{bmatrix}
\]  

\[G_v(i,j) = Y(i,j) \otimes H_v\]

Where \(Y(i,j)\) = luminous component of pixel \((i,j)\)

or

\[G_v(i,j) = -1 \times x(i-1,j-1) - 2 \times x(i-1,j) - 1 \times x(i-1,j+1) + 0 \times x(i,j-1) + 0 \times x(i,j) + 0 \times x(i,j+1) + 1 \times x(i+1,j-1) + 2 \times x(i+1,j) + 1 \times x(i+1,j+1)\]  \((2.6)\)

**Horizontal Convolution**

\[
H_h = \begin{bmatrix}
-1 & 0 & 1 \\
-2 & 0 & 2 \\
-1 & 0 & 1
\end{bmatrix}
\]

\[G_h(i,j) = Y(i,j) \otimes H_h\]

Where \(Y(i,j)\) = luminous component of pixel \((i,j)\)

or

\[G_h(i,j) = -1 \times x(i-1,j-1) + 0 \times x(i-1,j) + 1 \times x(i-1,j+1) - 2 \times x(i,j-1) + 0 \times x(i,j) + 2 \times x(i,j+1) - 1 \times x(i+1,j-1) + 0 \times x(i+1,j) + 1 \times x(i+1,j+1)\]  \((2.7)\)

The results from these convolutions are called vertical and horizontal spatial information of the video frame. Hence, the output of the Sobel filtered image at the \(i^{th}\) row and \(j^{th}\) column is given as:

\[y(i,j) = \sqrt{[G_v(i,j)]^2 + [G_h(i,j)]^2}\]  \((2.8)\)

This can be interpreted as the indication of the added or lost edges in the destination scene compared to the reference scene. Added edges result from impairments such as tiling, error blocks, and noise. Lost edges can result from impairments such as blurring.
Some statistical information about the spatial information can be obtained by computing the mean, variance, standard deviation and rms of $SI(i,j,n)$.

$$SI_{\text{mean}}(n) = \frac{1}{P} \sum_i \sum_j SI_r(i, j, n)$$  \hspace{1cm} (2.9)

$$SI_{\text{var}}(n) = \frac{1}{P} \sum_i \sum_j (SI_r(i, j, n) - SI_{\text{mean}}(n))^2$$ \hspace{1cm} (2.10)

$$SI_{\text{stdv}}(n) = \sqrt{SI_{\text{var}}(n)}$$ \hspace{1cm} (2.11)

$$SI_{\text{rms}}(n) = \sqrt{SI_{\text{var}}(n) + SI_{\text{mean}}^2(n)}$$ \hspace{1cm} (2.12)

Where:

- $n$ = Frame number
- $P$ = Total number of pixels
- $r$ = reference

**Temporal Information**

ITU-T P910 (09/99) has recommended the use of Temporal Information, $M_n(i,j)$ to describe the difference (movements) between two adjacent frames, $F_{n-1}(i,j)$ and $F_n(i,j)$ [ITU-T Recommendation P.910 1999].

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$$  \hspace{1cm} (2.13)

where $F_n(i, j)$ is the pixel at the $i^{th}$ row and $j^{th}$ column of $n^{th}$ frame in time.

It can be interpreted as an indication of the added or lost motion in the estimation scene compared to the reference scene. Added motion results from impairments such as jerkiness, error blocks and noise. Frame repetition introduces lost motion [Wolf et. al 1997].

As with the Spatial Information feature, valuable information can be obtained by computing the mean, variance, standard deviation and rms of $TI(i,j,n)$.

$$TI_{\text{mean}}(n) = \frac{1}{P} \sum_i \sum_j TI_r(i, j, n)$$  \hspace{1cm} (2.14)

$$TI_{\text{var}}(n) = \frac{1}{P} \sum_i \sum_j (TI_r(i, j, n) - TI_{\text{mean}}(n))^2$$ \hspace{1cm} (2.15)

$$TI_{\text{stdv}}(n) = \sqrt{TI_{\text{var}}(n)}$$ \hspace{1cm} (2.16)

$$TI_{\text{rms}}(n) = \sqrt{TI_{\text{var}}(n) + TI_{\text{mean}}^2(n)}$$ \hspace{1cm} (2.17)
2.7.d Metrics Based on Human Visual System

The ultimate aim of an objective video quality metric is to produce a model that will have a consistent and strong correlation with the subjective tests. Given the complexity of human perceptual system, that is very hard to achieve, and several techniques that claim to exploit the human-eye properties have been developed [Westen et al. 1995, Wu et al. 1996].

A general objective model should take into account the various aspects of visual information processing in the brain. These aspects include [Mullen 1985, Westen et al. 1995, Wu et al. 1996, Lindh and van den Branden Lambrecht 1996, van den Branden Lambrecht 1996, Winkler 1999]:

- Content
- Colour perception
- Spatial and temporal contrast sensitivity
- Pattern masking
- Viewing distance
- Display size
- Resolution
- Brightness
- Natural content versus synthetic

The matrices take into account that human visual perception are based less on absolute (luminance) values and more on contrast [van den Branden Lambrecht 1996, Mullen 1985], since the contrast sensitivity is much higher for luminance than for chrominance. The human visual system (HVS) is also less sensitive to high spatial frequencies as compared to low frequencies [Westen et al. 1995, Clarke 1995]. Hence, the overall colour and intensity of an image is more important than the very fine details [Winkler 1999].

Colour perception of the human eye is quite complex. Due to the anatomy and physical characteristics of it, humans are more sensitive to the colour red and less to yellow and blue. Details lost in those (yellow, blue) dimensions are therefore less evident and should account less in the video distortion. The same applies to the sensitivity to contrast alterations [Winkler 1999].
The developers of the matrices based on a complex spatial and temporal model of the human visual system (HVS) claim that the popular simple distortion measures, such as peak signal-to-noise ratio (PSNR) and root mean-square error (RMSE), do not correlate well with the perceived quality. This is because these simple tests do not take into account the viewing conditions and the psychological and physical behaviour of the human visual system. While claiming the superiority of these matrices, they concede that their evaluation is also more complex [Webster et al. 1993, Bhaskaran and Konstantinides 1997, Winkler 1999].

For a metrics to be widely acceptable it should [Wolf and Pinson 1999]:

- Emulate the subjective test results.
- Be suitable over a wide range of video sequences and codings
- Be computationally efficient

In 2000 the Video Quality Experts Group (VQEG) presented a report on the comparative study of ten commonly used objective tests and their performance against subjective tests. The metrics that were studied by VQEG were:

1. Centro de Pesquisa e Desenvolvimento – Image Evaluation based on Segmentation (CPqD-IES, Brazil)
2. NHK/Mitsubishi Electric Corp (Japan)
3. KDD (Japan)
4. Tektronix/Sarnoff (USA)
5. Ecole Polytechnique Fédérale Lausanne (EPFL, Switzerland)
6. TAPESTRIES (Europe)
7. National Aeronautics and Space Administration (NASA, USA)
8. Royal PTT Netherlands/Swisscom CT (KPN/Swisscom CT, The Netherlands)
9. National Telecommunications and Information Administration (NTIA, USA)
10. Institut für Nachrichtentechnik (IFN, Germany)

Surprisingly, their impartial and extensive evaluation of the matrices did not show significant superiority to the matrices over the simple objective tests. To summarise their findings [VQEG 2000]:

67
In spite of the advancements in the understanding of the human visual system and its application in the development of the metrics, no objective metric is able to fully replace subjective testing.

No objective metric statistically performs better than others in all reference conditions. They all have their own strengths and weaknesses; hence, are best used under restricted test conditions.

Surprisingly, no objective metric statistically outperforms the simple PSNR in all reference conditions.

Due to these findings, no single method can be recommended by VQEG to the ITU.

2.7.e Feng Xiao’s DCT-based VQM

Figure 2.26 gives the correlation between the Spatial Frequency, Temporal Frequency and Contrast Sensitivity of an eye and shows that the spatiotemporal contrast sensitivity of eye to changes in the rate of flicker and resolution increases and then decreases. Feng Xiao has developed a simple DCT based quality test based on this human spatial-temporal contrast sensitivity function for testing the quality of MPEG4 videos [Xiao 2000]. The program for the calculation of Feng Xiao’s DCT-based VQM is freely available [Video Quality Studio 2006].

2.7.f Weaknesses of Video Quality Assessment Techniques

A number of subjective and objective test methods have been developed to measure the quality of digital video, and these methods are very helpful for the evaluation of digital video components (video codecs, monitoring transmission quality, understanding psychophysical aspects of quality), but they still suffer from a number of limitations when applied for testing the quality of multimedia streamed over the Internet. Some of these limitations are [Miras 2002]:

- Most of the tests are designed for testing short video sequences (typically approximately 10 seconds duration), but these 10-sec video sequences are not long enough to experience all kinds of impairments that would occur in a real Internet video application. This problem can be partially tackled by employing continuous assessment techniques.

- The tests developed were for the quality assessment of broadcast / cable TV transmission, and not particularly for transmission over the Internet. As a result, some of the artefacts of Internet transmission have not been considered. For example, since TV broadcasts can provide an assured transmission channel, with bounded delay the objective models can handle the coding artefacts, but when it comes to Internet transmission they are not designed to track distortions due to large delays and jitters.

- The tests are designed solely for the video, i.e. in all the cases the audio is kept mute. It has been irrefutably shown that the quality of video "experienced" by viewers is affected by other non-visual conditions, such as the quality of audio and lip synchronisation.

2.7.g Summary of the Literature Review for this Section

From the literature survey, the following points stand out:

1. Both subjective and objective tests should be carried out to evaluate the quality of video sequences.

2. Since all objective tests work well within limited premises, the videos should be tested using a number of objective test methods. Results of a single objective test method may not be enough to make proper deductions.

3. Simple objective tests, PSNR, SI and TI are often as good as more complicated objective tests based on human visual system. The limitations of the tests should be kept in mind while drawing conclusions from the experimental results.

4. Feng Xiao’s DCT-based VQM is a freely available tool that can be used for objective evaluation of a video sequence [Video Quality Studio 2006].

69
CHAPTER 3

PROPOSED ARCHITECTURE
CHAPTER 3
PROPOSED ARCHITECTURE

3.1 INTRODUCTION

The main idea of the work is to understand the human psycho-physical visual system and use this understanding to transmit an e-learning video in the best possible perceptual quality within the constraints of the available bandwidth.

As explained in Chapter 2 of this thesis, there are a number of ways to efficiently transmit video over the Internet, but most of these methods only consider the requirements of the network. In this research it has been proposed that it is possible to improve the quality of transmitted video by first reducing the amount of data that needs to be transmitted and then by transmitting the essential data in a way that it maximises the user-perceived quality. In this research it is intended to achieve this by bridging the gap between the requirements of the network and the psycho-physical aspects of the human visual system.

This chapter describes the requirements for transmitting an e-learning video over the Internet and the proposed solutions.

3.2 REQUIREMENTS FOR THE TRANSMISSION OF AN E-LEARNING VIDEO OVER THE INTERNET

Students attending a classroom lecture experience a rich visual experience that is simultaneous, detailed and coherent [Rensink 2000a]. That is, if the students take a panoramic view of the class, they see a number of events at the same time, in high resolution (or high density of information) and all the sections of the image are combined in a proper jigsaw puzzle. The students would like to experience the e-learning video as close to the actual classroom experience as possible [Mukhopadhyay and Smith 1999]. Unfortunately, the students are often linked with very limited bandwidth of up to 60 kbps. On top of that, the e-learning tools have to transmit audio and data files to the students. Hence, the available bandwidth gets partitioned among many applications. Since the sender can only transmit at a rate equal to that of the bottleneck bandwidth link, the challenge is to improve the quality within that limitation.
On the other hand, while delivering a lecture an instructor has to continuously work hard to hold the interest of students. This issue becomes even more important in the case of e-learning where the options available to the instructor are limited and he is forced to present it sitting in front of the camera with only a head-and-shoulder view being transmitted.

When the developers of e-learning tools look at the surroundings, they “feel” as if they see a detailed coherent world, and then they try to represent this in the videos. Unfortunately, due to the limitations set by the availability of the bandwidth and other recourses, there is a big difference between the quality of video that is expected by the viewers and that provided to them.

The understanding of the human visual system can help in bridging the gap between the expectation and the reality by suggesting ways to efficiently transmit the video so as to provide the best perceptual quality within the limitations of the network. For example, at a particular instance, humans only see either the gist of the panorama or the detailed view of a small part of it. At no instance of time do they have detailed representation of the whole scene. This weakness of the visual system may be exploited for improving the transmission of video.

For many years film-makers have informally applied the knowledge of these limitations of human perception to make “continuous” and engaging movies. People watching those films do not complain about the limitations of the perception, on the contrary, they often find the presentation much more appealing than real life. If that is the case, then the reason for dull e-learning video lectures is not the limitation due to the use of video for teaching, but the actual limitation arises due to the style of presentation. Hence, efforts should be directed towards making a video presentation according to the principles by which humans perceive the real world. By applying the knowledge of strengths and weaknesses of human perception it may be possible to make the lectures more exciting and captivating.

Unfortunately, not many instructors are trained to be good presenters and video-graphers. Most often an instructor just sits in front of a monitor with a camera placed over it and delivers the whole lecture without any movement. Unfortunately this mere capturing of a talking-head-and-shoulder video does not make an interesting video. More interesting than this is a “real lecture” video, where the lecturer has the freedom to move around, use an electronic whiteboard and show real models as he / she would do in a real classroom environment [Mukhopadhyay and Smith 1999].
Thus the transmission of an e-learning video over the Internet has its own specific requirements that need to be satisfied for obtaining best results. Some important requirements are:

**Requirements of E-Learning**

- Transmission within low bandwidth, since, many students are connected by ADSL link.
- Sharing of bandwidth with other applications such as audio, data transmission, etc.
- Live transmission of video in synchronous e-learning.
- Transmission of video in a way as to excite interest among the viewers, i.e., shots taken from different angles and zooms.

**Requirements of the Viewers**

- Very low degree of packet delays, jitters and losses.
- Sharp image quality.
- Smooth (non-jerky) movement of the video.

### 3.3 SET-UP FOR A VIRTUAL CLASS ROOM

Unfortunately, it is not possible for every instructor to be a good film-maker. There is therefore a need for automating the “art” of film-making and this is possible only within a limited domain. The first important step to make the presentation captivating is to break out of the head-and-shoulder mould and present the lecture in a “real” classroom setting. In this condition, a multiple camera system can take shots from different angles and using different zooms. A combination of fixed and Pan/Tilt/Zoom (PTZ) cameras can cover the lecture dais quite effectively [Mukhopadhyay and Smith 1999].

In this section the proposed set-up for recording the video of lectures has been described. In order to take shots from different angles and views, the challenging task is the placement and automatic switching of digital video cameras [Mukhopadhyay and Smith 1999]. In the system depicted in Figure 3.1, the video shots can be taken using three different cameras – one fixed and two tracking PTZ (Pan/Tilt/Zoom) cameras. Various techniques have been proposed to switch between the cameras, in this section the hardware required for switching shall not be considered, but it shall be assumed that an appropriate switching method is available.
Camera 1. A tracking PTZ camera to take a close-up shot of the instructor's head and shoulders.

Camera 2. A tracking PTZ camera to take mid-shot so as to show the instructor and the region of interest next to the instructor, for example, the whiteboard, model, etc.

Camera 3. A fixed camera covering the panoramic view which shows the entire lecture dais from which the instructor presents his lectures.

The placement of cameras is an important issue. Camera 1 takes close-up shots and has the responsibility to show that the instructor is maintaining eye contact with the students. This can be achieved by "direct gaze technique", which assumes that a remote viewer cannot distinguish whether the instructor is looking at them or not as long as the instructor looks towards the camera within a certain angle. This angle has been found experimentally to be 2.7 degrees horizontally and 9 degrees vertically [Chen 2001, Chen 2002]. Hence, while speaking the instructor should look towards the Camera 1 as much as possible. Camera 2 should be placed at a position that will provide the best view of the object of interest. A reasonably good placement of Camera 3 is at the back, from where it can take clean panoramic shots.
The use of these three cameras for giving different angles/areas of view may give the students a sense of presence in real life, and create an illusion that the video has been edited by hand [Mukhopadhyay and Smith 1999]. This makes the output video more interesting than that produced by using a single camera to capture head-and-shoulder video.

Based on the same concept and depending upon the requirement and availability of resources, more than three cameras may be used.

3.4 OPTIMUM TRANSMISSION OF VIDEO

Having captured video from different angles and zooms, this needs to be efficiently transmitted over the network. The main issue in this is the transmission of different types of video sequences, i.e., video sequences taken from different angles and zooms, so as to provide the best perceptual quality to the viewer.

In this research it has been proposed that it is possible to significantly improve the quality of video at the receiver end by first reducing the amount of data to be transmitted, and then by transmitting the important data in a controlled way so as to provide the best perceptual quality.

In the first part of the research, it has been proposed that the removal of background from the video sequence would help in optimal utilisation of the bandwidth. An extensive amount of work has been done in separation of foreground and background. This feature is required for various purposes such as, codec compression, tracking of moving object, face recognition, security surveillance, film-making, etc. One way the MPEG defines background is using the concept of conditional replenishment. As discussed in Section 2.4.a, in the conditional replenishment the “background” is the portion of the frame that witnesses change less than the set threshold for a certain duration. In spite of all the research work carried out in this direction, there is still a need to find better ways of identifying “background” from “foreground”, because not much work has been done on separation of “foreground” and “background” based on the principles of psychology. It has been proposed here that this identification can be guided by the understanding of the human visual system.

Once the foreground has been separated from the frame it needs to be transmitted over the network. Under normal conditions the fluctuation in the quality of video is mostly due to the random variations in the network conditions, but in the proposed investigation the use of multiple cameras will introduce additional fluctuations due to the sudden changes in video scenes. Hence, there is a need for a dynamic adaptive procedure to provide consistently good quality video in real-time.
Chapter 3

Proposed Architecture

A considerable amount of work has also been done on adaptation of video to suit the fluctuations in network. Most of these approaches only consider the network for optimisation and relatively less work has been done on adaptation of video based both on the network conditions and perceptual quality of video.

3.5 SEPARATION OF BACKGROUND AND FOREGROUND

One way of streaming the video at high perceived quality, while at the same time saving on the bandwidth, is by transmitting the relevant portions of a frame at a high resolution and the background at a relatively lower resolution, or infrequently. In MPEG-4 this is achieved by coding in terms of the Video Objects Planes (VOPs), wherein different sections of the frames can be encoded separately and sent in layers. For example, in a news broadcast, the news reader can be encoded as a separate foreground video object and the background as another video object (Puri and Chen 2000).

While MPEG-4 has the provision for transmitting the foreground and background in different layers, it does not recommend any particular method of separating the two. Hence, in recent years, separation of foreground from background in natural video has attracted considerable interest and a number of schemes have been developed in order to do that.

A separation scheme involves two steps:

- Identification of the portions comprising the foreground in a frame; and
- Separation of the foreground from the background.

For the identification of the foreground from background, the human visual system works in a very complex way and to date, this understanding is incomplete. As discussed briefly in Chapter 2, human eyes distinguish between the background and foreground on a case-to-case basis. For example, in Figure 3.2a, the sunflower is a foreground, but in Figure 3.2b, the same sunflower may become a background, provided the observer is interested in the moth.

Once a proper basis for identification of foreground has been identified, suitable algorithm(s) for separation have to be adopted. The separation techniques may be based on spatial and temporal differences, such as relative motion, gray scale luminosity, colour scheme, etc., between the two.
In a hypothetical classroom, illustrated in Figure 3.3, there are a number of objects. Some of these objects are stationary, such as blackboard, table, ball-and-stick model, envelope, paperpunch, paper shredder; and some of them are moving, example instructor, fan, clock. A common way of separating foreground from background objects is to identify and separate the "moving objects" from the "stationary background". Hence, if the foreground is identified as the moving objects, then instructor, fan and the hands of the clock would be identified as foreground, and the others as the background. Here it is pertinent to ask questions like:

- Is the moving fan a foreground?
- Is the stationary ball-and-stick model which the instructor is discussing, a background?
- If the instructor, a moving object, and the model, a stationary object, have been identified as foreground objects, how can they together be separated from the rest of the frame?
- Once the foreground and the background have been transmitted separately, how good would the video be after re-combination at the receiver end?
- Will the viewers be able to notice slight discrepancies that will arise due to this technique?

In this research work these questions have been addressed.
3.6 ADAPTATION OF VIDEO TRANSMISSION USING FEED-FORWARD CONTROLLERS

Figure 3.1 shows schematic representation of the proposed architecture. In this, the video streams from the cameras are sent to the codec and from there they are transmitted to the Internet. These video streams are also sent to the Feed-Forward Controller where the sequences are dynamically characterised. Depending upon the characteristics of the sequence and the availability of the bandwidth, the controller calculates and adjusts the transmission of the video stream so as to provide the best perceptual quality to the viewers. For this research two different schemes have been investigated.

3.6.a Use of Multi-Layer Feed-Forward Controller for Optimised Video Transmission

In the first scheme, the optimised transmission of video in a heterogeneous environment has been studied. In a synchronous e-learning class the instructor and students are often connected in a heterogeneous environment, where most often the instructor is connected to the network by a large bandwidth and the students by relatively smaller bandwidths. The transmission of video in such a heterogeneous environment becomes a challenging task because when the video is transmitted at the bottleneck rate, the large bandwidth connections are incompletely utilised. On the other hand, when the transmission rate is adjusted according to the capabilities of the large
bandwidth connections; the receivers connected by smaller bandwidth link experience more packet delays and drops. In either of the two conditions, the video quality is suboptimal.

Layering of the video stream has been proposed as a technique for the reliable multicast of video within a heterogeneous environment. In this technique, the most important data is sent in the base layer that can be accepted by all receivers, and the rest is in the enhancement layers. At the receiving end, these layers can be combined, depending on the bandwidth limitation of the receivers to develop the program. The quality of the program depends upon the number of layers combined.

In this research work it has been proposed that there is a need for a good layering technique for the transmission of video. For this three aspects need to be considered:

- The number of layers the video should be divided into.
- The distribution of available bandwidth among the layers.
- Frame rate of the transmission.

In order to optimally partition the video stream into layers, the use of a Multi-Layer Feed-Forward Controller has been researched. This controller: (a) dynamically characterizes the nature of video sequence; (b) divides the video sequence into layers in an optimal way before transmitting it into the network. The proposed architecture is for the optimal post-codec temporal layering of video in Intra H.261 codec. With modifications, it can be made more general and used for other codecs also.

### 3.6.1 Use of Single-Layer Feed-Forward Controller for Optimised Video Transmission

The current network does not support layering schemes, hence, it was decided to study the optimization of video transmission in one layer. In this research project it has been proposed that the available bandwidth can be most efficiently utilised by dynamically adapting the transmission of video, based on the availability of the bandwidth. Here the adaptation of the video stream is done within a two dimensional adaptation space defined by frame rate and spatial resolution. These encoding variables have been chosen as they most closely map to the spatial and temporal complexities (action and detail) of the video content. In this research work, use of a Single-Layer Feed-Forward Controller for dynamic adjustment of the transmitted video stream has been investigated. The controller (a) dynamically characterises the ongoing video stream, and (b) adjusts the frame rate and resolution so as to provide the observers perceptually optimised video within the available bandwidth. While the current work has been
carried out using MPEG-4 codec but with modifications, it can be made more general and used with other codecs as well.

3.7 ADAPTATION OF VIDEO TRANSMISSION USING MULTI-LAYER FEED-FORWARD CONTROLLER

Depending upon the characteristics of the sequence and the availability of the bandwidth, the Multi-Layer Feed-Forward Controller calculates (a) the number of frames, (b) distribution of frames among different layers, and (c) frame rate that would give the best perceptual quality to the viewer within the available bandwidth.

In this controller the independent variables are:

(a) Complexity of the video sequence – defined by the number of Macroblocks in a frame,
(b) Distribution of frames among different layers,
(c) Number of layers,
(d) Frame rate,

The dependent variables are

(a) Bandwidth requirement,
(b) Perceptual quality of video; and

1. Complexity of Video Sequence. The complexity of a frame has been defined as the number of Macroblocks sent per frame. This is because the architecture of the Intra H.261 is such that the codec sends only the changed blocks, hence, a sequence with less motion has less Macroblocks per frame and a sequence with high motion has more Macroblocks per frame. Hence, the number of Macroblocks sent per frame depends on the number of changes in the scene. This increase in complexity results in an increase in the load on the sender and on the network.

2. Distribution of frames among layers. The distribution of frames among different layers is an important parameter. It will depend upon the available bandwidth of the different receivers. The base layer normally contains the most useful information and ideally the major share of the bandwidth should be allocated to it. On the other hand, if a large part of the bandwidth is given to one layer, the user would not have a strong
reason to subscribe to enhancement layers. Hence, the information must be judiciously divided among the layers.

In this research for distributing the frames between the layers two algorithms were tested (Figure 3.4):

Algorithm A: Using this algorithm the successive frames are sent equally on the different layers. To calculate the layer number to which a frame would go, the following equations are used:

\[ R = \lfloor n \mod (M+1) \rfloor \]

\[ L_m(n) = M + 1 \text{ if } R = 0 \text{ else } L_m(n) = R \]

Where

- \( R \) is remainder of the block update during frame time \( n \) and the maximum number of layers \( (M+1) \),
- \( n \) is the frame time,
- \( M+1 \) are the maximum number of layers,
- \( L_m(n) \) is the layer number during frame time \( n \).

Algorithm B: In this algorithm the frames were sent on different layers so that more frames were sent to higher numbered layers [McCann 1996]. In this scheme the base layer requires very less bandwidth but is able to take only the most essential data.

\[ N_L \text{ (No. of frames in a layer } L) \propto 2^{L-1} \]

Hence, a layer \( L_x \) had twice the number of frames sent on layer \( L_{x-1} \).

\[ L_m(n) = M - r(n \mod 2^M + 2^M) + 1 \]

\[ r(n) = \min\{k > 0 : \lfloor n/2^k \rfloor 2^k = n \} - 1 \]

Where

- \( L_m(n) \) is the layer number during frame time \( n \),
- \( n \) is the frame time,
- \( M+1 \) are the maximum number of layers,
- \( r(n) \) is the bit position of the right-most non-zero bit in the binary representation of \( n \),
- \( k \) is any layer with \( k > 1 \)
FIGURE 3.4. Distribution of frames on different layers for Algorithms A and B [Thakur and Motyckova-Carr 2003].

3. **Number of Layers.** With the increase in the number of layers there is an improvement in performance due to:

- Ability to handle greater heterogeneity because a finer division of layers allows the heterogeneity to be handled better.
- The finer division of layers allows the receiver to adjust its reception nearer to the available bandwidth thereby leading to an improvement in the video quality due to optimum utilization of bandwidth.

On the other hand it increases the problems due to:

- Increase in the complexity of the codec resulting in greater computational load on the sender. **Figure 3.5** shows the results of the experiments carried out to study the effect of number of layers sent on the load on the sender. The figure shows an increase in the computational load on the sender with an increase in the number of layers sent. This increase in the computational load on the sender depends upon the way the layering is carried out. **Figure 3.6** shows the effect of two different algorithms on the load on sender. The figure shows that the simpler algorithm, Algorithm A, requires less computational power compared to Algorithm B. Hence, the computational load on the sender needs to be calculated based on a number of factors like the number of layers and the algorithm used.
Decrease in the codec efficiency resulting in a wastage of bandwidth because the header of each packet has also to include the layer number. Based on the literature survey and the results of the preliminary experiments a generalised diagram (Figure 3.7) has been developed [Thakur and Motyckova-Carr 2001]. It shows the effect of the number of layers on various parameters.

Therefore, increasing the number of layers follows the law of diminishing returns. Beyond a certain number, any further increase in the number of layers only makes the overall quality worse. This optimum number depends upon the type of codec and the partitioning method. In this work a maximum of 6 layers have been tested since it has been assumed that 6 layers are sufficient for the proper transmission.

4. **Frame rate**: The frame rate is an important parameter. As the frame rate increases the perceptual quality of video also increases, but this happens at the expense of increased bandwidth requirement and load on sender. In the present work the frame rate has been manually controlled but in future it shall be done automatically.

**FIGURE 3.5.** Effect of number of layers on the load on sender (read from the percentage usage of CPU in Windows Task Manager).

(Algorithm A, 3 fps, 9 macroblocks / frame)

**FIGURE 3.6.** Effect of number of layers on the load on sender for a movie sequence.
5. **Bandwidth Requirement:** For a video sequence the bandwidth requirement increases with the increase in the number of layers due to an increase in overheads, hence, the number of layers should be minimum [Handley and Jacobson 1998].

6. **Perceptual Quality of Video:** The perceptual quality of video increases with the increase in the number of layers joined. The initial investigations have shown that even if all the layers are joined the quality deteriorates with the increase in the number of layers. Hence, the number of layers should be kept to the minimum [Handley and Jacobson 1998, Handley et al. 1999].

### 3.8 ADAPTATION OF VIDEO TRANSMISSION USING SINGLE-LAYER FEED-FORWARD CONTROLLER

Depending upon the characteristics of the sequence and the availability of the bandwidth, the controller calculates the frame rate – Resolution combination which would give the best perceptual quality to the viewer.

In this controller the dependent variables are

(a) Bandwidth requirement,

(b) Perceptual quality of video; and
The independent variables are:

(a) Complexity of the video sequence – defined by the Spatial Information and Temporal Information of video sequence,
(b) Frame rate,
(c) Resolution.

1. **Bandwidth Requirement.** Bit rate of the video sequence is dependent upon the size of frame and frame rate.

\[
\text{Bandwidth req.} = \text{Frame Height} \times \text{Frame Width} \times \text{Frame rate} \times \text{Codec Compression Ratio} \tag{3.5}
\]

Experiments have shown that for a subject to recognise emotions on the face, the image should have a minimum visual angle of 6 degrees and ideal of 14 degrees. For a student sitting about half a meter from his monitor, this approximately corresponds to an image size of 50-125 mm (200-500 pixels) wide, hence the size should be fixed in this range [Chen 2001]. For a smooth motion the viewer would like to have frame rate as high as possible and up to 25 fps.

2. **Perceptual Quality of Video.** The perceptual quality of video increases with an increase in the resolution, which in turn is dependent upon the frame size, frame rate and the codec type and setting. For example, at a fixed bandwidth, if there is an increase in motion the video will become smoother by increasing the frame rate, decreasing the resolution and decreasing the uncompressed key frame (I frame in MPEG-4) rate.

3. **Spatial Information (SI) of Video Sequence.** It is a measure that generally indicates the amount of spatial detail of a picture. It is usually higher for more spatially complex scenes. It is figured by calculating the spatial edge noise, which are the changes around the edges due to the spatially varying distortion in close proximity to the edges of the objects. It plays an important role in image quality. It is interpreted by calculating the added or lost edges in the destination scene compared to the source scene. Added edges result from impairments such as tiling, error blocks, and noise. Lost edges can result from impairments such as blurring. Since MPEG-4 sends the changed macroblocks, any change in edges will be sent, hence an increase in Spatial Information will result in an increase in bandwidth requirement.

4. **Temporal Information (TI) of Video Sequence.** It is a measure that generally indicates the amount of temporal change of a video sequence. It is usually higher for high motion sequences. It describes the difference (movement) between two adjacent frames and is
calculated based on temporal information as indicating added or lost motion in the estimation scene compared to the source scene. Added motion results from impairments such as jerkiness, error blocks, and noise. Like the Spatial Information, the increase in Temporal Information will result in an increase in bandwidth requirement.

5. Frame rate. The frame rate is the number of frames or images that are projected or displayed per second. Higher frame rates produce smoother motion in a video sequence, but it also results in higher bandwidth requirements.

6. Resolution. Visual resolution is the smallest detail that can be seen and is characterised by the number of pixels packed on the screen. Naturally subjects like to have high resolution but that comes at a high cost in terms of bandwidth and the computational power. The higher the visual resolution, the greater the bandwidth requirement. Also with an increase in resolution, the amount of work the video card must do in order to create a frame of data increases exponentially.

3.9 CODECS USED FOR THE RESEARCH

3.9.a Intra H.261

One of the off-shoots of H.261 is the Intra-H.261 codec. This codec was originally made for vic (vic - is a video conferencing tool developed at University of California at Berkeley). Since it is now freely available it has also been used in various commercial video conferencing tools [McCanne 1996].

The frames in MPEG, H.261, H.263 codec mainly consist of Intra (I), Bi-directional (B), Predicted (P) frames. The first frame, the I frame, is not predicted in anyway. The other – Bi-directional and Predicted – frames are formed by predicting the previous frames. The Intra-H.261 codec is very closely related to the H.261 codec but sends only the I frames. Hence, the algorithm becomes simpler in terms of complexity as it does not have to compute the prediction error signal. In spite of its relative simplicity it provides the benefit of aggressive conditional replenishment [McCanne 1996].

This codec was selected for the work because of its free availability and relative simplicity of use.
3.9.b MPEG-4

A number of MPEG-4 codecs are commercially available, but in this research project a freely available codec – FFmpeg – was used. FFmpeg is an open source codec and has the following components [FFmpeg Multimedia Systems 2006]:

- ffmpeg – a command line tool to convert one video format to another.
- ffserver – a HTTP multimedia streaming server for live broadcasts.
- ffplay – a simple media player based on the ffmpeg libraries.
- libavcodec – a library containing all the ffmpeg audio / video encoders and decoders.
- libavformat – a library containing parsers and generators for all common audio/video formats.

This codec suffers from some limitations like:

- It is not well-documented.
- It does not support VOPs.

In spite of these limitations it was selected due to its free availability.

3.10 PREPARATION OF TEST VIDEO SEQUENCES

3.10.a For Multi-Layer Feed-Forward Controller

Since the type of video sequence has significant effect on the compression scheme of the codec, eight different types of video schemes were transmitted and tested. These video sequences were recorded in a VCR and played when required. Their complexity was judged by counting the average number of Macroblocks per frame. The simplest one was the image of a stationary object transmitted continuously, it had a complexity of one Macroblock per frame. At the other extreme was an image in which the scene was continuously zoomed in and out, having a complexity of 14 Macroblocks per frame. In between these two limits were different kinds of video sequences, like people talking, sports, soldiers marching, conferencing.

3.10.b For Single-Layer Feed-Forward Controller

Video Quality Experts Group (VQEG) has provided twenty standardised test sequences. Initially these test sequences were used for the development of the equations for the controller. These sequences had wide variations in the values of SI and TI, hence, were not found to be suitable for the development of equations for the restricted scenario of the classroom (Appendix 2).
FIGURE 3.8. Stills from the representative video sequences recorded for developing the controller.

Since the study is restricted to the development of a tool for application in the transmission of e-learning classes, the sequences required were restricted to the classroom scenario. The use of multiple cameras was simulated by recording ten test sequences from different angles and zooms. Of these ten sequences three shots were taken of the head-and-shoulder sequences, three of the middle sequences and four of the panorama sequences. Figure 3.8 shows stills from the three representative video sequences.

The recordings were carried out in an actual classroom. The major source of light was from overhead florescent tubes but some light also came in through the blind-covered windows. As a result there was a gradual change in the light condition in the classroom.
3.11 CHARACTERISATION OF VIDEO SEQUENCES

The need to measure video quality objectively arises while performing video and image processing for various multimedia applications. The objective measurement tests vary from a simple difference between original and distorted sequences to very complicated tests that are based on Human Vision System (HVS) models and include complex mathematical calculations. Both the original video and the degraded video are digitised and processed to extract a large number of features. The processing may include Sobel filtering, Laplace filtering, fast Fourier transforms, first-order differencing, colour distortion measurements, and moment calculations.

3.11.a For Multi-Layered Feed Forward Controller

In this study, the aim has been to divide the load on the bandwidth, into different layers. Intra H.261 uses temporal compression model called conditional replenishment, wherein a video frame is made up of blocks of $8 \times 8$ or $16 \times 16$ pixels and only those blocks that change substantially in successive frames are encoded and sent. As a result the load on the bandwidth is proportional to the number of Macroblocks sent, and the number of Macroblocks sent itself depends upon the changes that take place in the video sequence. Hence, the number of Macroblocks transmitted can be considered as a measure for the complexity of the video sequence: higher the complexity – changes in the video sequence – higher is the number of Macroblocks transmitted.

Figure 3.9 shows the experimental values for the relationship between the complexity (number of Macroblocks per frame) and the bandwidth requirement. The figure shows almost a linear relationship between the two. Since the value of complexity will be used to control the division of bandwidth based on the requirements of the video sequence and there is a strong correlation between the number of Macroblocks per frame and the bandwidth requirement, hence it is justified to use this definition for characterization of the video sequence.

3.11.b For Single-Layered Feed-Forward Controller

In this study, individual images of a video sequence need to be characterised in order to characterise the type of ongoing video scene. While ITU-T P910 (09/99) has recommended the use of Spatial Information (SI) and Temporal Information (TI) to calculate the objective quality of the video sequences with respect to the reference sequences (Section 2.7.c), the same calculation can be adapted to study the differences between two successive frames. In the development of the controller, the need is to calculate the difference between two successive images, and not to find how “pleasing” the video seems to viewers. Due to this reason the SI and TI based evaluation method was selected for the development of the controller.
FIGURE 3.9. Effect of complexity on bandwidth requirement.

Since MPEG-4 uses conditional replenishment and therefore sends the difference between two frames, SI and TI are also indirectly related to the bandwidth requirement. Hence, they have been selected for the characterisation of the video sequences. The only divergence from the recommendations by ITU-T P910 (09/99) is that while ITU recommends taking the highest values of SI and TI of the whole sequence for reporting, in this project the individual values of the frames have been kept for further calculations by the controller.

**Spatial Information:** Figures 3.10.a, 3.11.a and 3.12.a show the Nth frame of the three test sequences and Figures 3.10.b, 3.11.b and 3.12.b show the (N+1)th frame of the three test sequences. Figures 3.10.c, 3.11.c and 3.12.c show the results of the edge detection of the Figures 3.10.a, 3.11.a and 3.12.a, using Sobel Filter. The figures show that the area covered by the non-white region is maximum for the Panorama sequence, followed by the Head-and-Shoulder and minimum for the Middle. This corresponds with the SI values of the three sequences, which are in the same order.

**Temporal Information:** Figure 3.10.d, 3.11.d and 3.12.d show the difference between the two consecutive frames of the three sequences. The figures show that the area occupied by the black pixels is maximum for the Panorama sequence, followed by the Head-and-Shoulder and minimum for the Middle. This corresponds with the TI values of the three sequences, which is in the same order.

From this preliminary study it seems that the SI and TI can be used to objectively characterise each frame.
Proposed Architecture

FIGURE 3.10. (a) and (b) Two consecutive frames of the Head-and-Shoulder sequence, (c) Edge detection of \( N \)th Frame using Sobel Filter, (d) Difference between the two frames.

Note: Figures (c) and (d) have been enhanced using "Negative" and "AutoBalance" modes of Microsoft Photo Editor.

Spatial Information of the Head-and-Shoulder Sequence = 52.6
Temporal Information of the Head-and-Shoulder Sequence = 4.4
3 Chapter. Proposed Architecture

FIGURE 3.11. (a) and (b) Two consecutive frames of the Middle sequence, (c) Edge detection of Nth Frame using Sobel Filter, (d) Difference between the two frames.

Note: Figures (c) and (d) have been enhanced using "Negative" and "AutoBalance" modes of Microsoft Photo Editor.

Spatial Information of the Middle Sequence = 38.8
Temporal Information of the Middle Sequence = 2.9
Sobel Filter of the Nth Frame of the Panorama Sequence
Difference between the (N+1)th and Nth Frames of the Panorama Sequence

FIGURE 3.12. (a) and (b) Two consecutive frames of the Panorama sequence, (c) Edge detection of Nth Frame using Sobel Filter, (d) Difference between the two frames.

Note: Figures (c) and (d) have been enhanced using "Negative" and "AutoBalance" modes of Microsoft Photo Editor.

Spatial Information of the Head-and-Shoulder Sequence = 55.1
Temporal Information of the Head-and-Shoulder Sequence = 6.3
### 3.12 SUMMARY AND CONCLUSIONS

This chapter discusses the requirements for transmitting an e-learning video over the Internet and then describes proposed architecture for streaming the e-learning video within the constraints of the available bandwidth.

Current research in human visual system has shown that the human eye can only see one section of the scene at a time — either in detail or in a coherent (stable) way. Humans cannot see more than one object in detail and in a coherent way at the same time; and also we can focus attention only on one event at a time [Rensink 2000a, Rensink 2000b]. It has been proposed that these limitations of human eyes be exploited to define “foreground” and “background”. Once
these two have been identified, they may be separated using suitable techniques. The foreground can then be transmitted in a perceptually optimised way within the available bandwidth.

In order to vary the quality of video according to the availability of bandwidth use of Feed-Forward Controllers has been proposed. In the first scheme, Multi-Layer Feed Forward Controller, the controller inputs the available bandwidth and characterises the ongoing sequence. Based on the number of Macroblocks per frame, the controller divides the video into a number of layers before transmitting it. In this scheme the optimisation is done based on the number of layers, division of frames among different layers and the frame rate. Such a scheme is suitable for satisfying the heterogeneous bandwidth connections in the network.

For carrying out the research work, Intra H.261 based videoconferencing tool was used. Eight different video sequences were taken and characterised for the number of Macroblocks per frame.

In the second scheme, Single-Layer Feed-Forward Controller, the controller inputs the video and dynamically characterises it regarding the Spatial Information and Temporal Information. Based on the video characteristics and the availability of bandwidth, the controller then adapts the video transmission in terms of frame rate and resolution so as to provide a perceptually optimised video. Such a scheme is suitable for adjusting the quality of transmitted video so as to achieve the best perceptual quality within the constraints of the available bandwidth.

For carrying out the research work, FFMPEG has been identified as the MPEG-4 codec. Ten different test sequences have been prepared and characterised for their SI and TI values, based on the recommendations of the ITU-T P910 (09/99).
CHAPTER 4

HUMAN VISUAL SYSTEM BASED FOREGROUND-BACKGROUND SEPARATION
CHAPTER 4

HUMAN VISUAL SYSTEM BASED FOREGROUND-BACKGROUND SEPARATION

4.1 INTRODUCTION

One of the major goals of video transmission is to provide the highest possible perceived video quality to the receiver within the constraints of the existing network. In this research it has been proposed that it may be possible to achieve this by first efficiently removing the redundant background data from the video and then transmitting the important foreground data in an efficient way. Since there is no clear description for the “redundant background” and “important foreground”, this chapter describes the research carried out to understand how the human visual system distinguishes between foreground and background objects in an e-learning video. The insights gained from the results of the psychophysical experiments have been used to propose schemes for efficient separation of foreground objects from the e-learning video. This work forms the basis for the next chapter, which describes the study carried out to transmit the reduced video data efficiently within the available bandwidth.

This chapter discusses the interesting prospect for improving the transmission of an e-learning video by separating the foreground from the scene. Research has been done to understand the psychological aspects of the identification of foreground objects in a scene. Based on the results, work has been carried out to combine some relevant foreground separation techniques in order to obtain the desired separation. Finally, the strengths and weaknesses of the proposed scheme have been identified and discussed.
4.2 OVERVIEW OF THE PSYCHOPHYSICAL EXPERIMENTS

As discussed earlier, the human visual system is a highly complex system which has not been fully understood; but the existing knowledge can still be exploited for improving the transmission of an e-learning video over the Internet.

Some of the relevant observations made by other researchers have been:

- **Physical Aspects of Eyes.** The light sensitive portion of the human eye is the retina, which is comprised of rod cells which respond to dim light and handle lower-resolution, black-and-white, night vision; and the cone cells, which are sensitive to bright light and handle high-resolution vision and colour vision. The peripheral retina is rod-dominated and the central retina is cone-dominated. In the centre there is a very small fovea that has the highest concentration of cone cells and hence is the most sensitive part of the eye. Therefore, the human eye can have a detailed view of objects only within a very narrow region, of about 4°, and any object out of this small window is seen at low resolution. Depending upon the requirements of the task, the human eye moves so as to maintain the most important portion of the image on the fovea [Cater et al. 2002, Cater et al. 2003b].

- **Perceptual Aspects of Eyes.** It is not possible for the human brain to process all the information sent by the eyes to the brain. The information the brain processes depends upon objects that "catch the eye", example, the moving object or brightly coloured object, during the bottom-up processing; or the task-at-hand during the top-down processing. In an e-learning class the students have to consciously pay attention to the instructor, hence, the visual attention process involved would fall under the regime of the top-down processing. In order to reduce the amount of information to be processed, subjects tend to ignore all the objects that have been identified as objects of low interest and will not take a detailed view of those objects, even if the eyes’ gaze moves over it. As a result, they will not be able to notice the characteristics of the background (Inattentional blindness) or remember the presence or absence of objects of low interest (Change blindness) [Cater et al. 2003a, Cater et al. 2003b].
One of the aims of the research areas is to separate foreground, objects of high interest, from
the background, objects of low interest, in an e-learning classroom video. Hence it is worth
investigating the following:

- What are the regions of high and low interest in an e-learning video?
- How do subjects respond to the spatial changes, i.e., the perceptual quality, of the
  foreground and background?
- How do subjects respond to the temporal changes in the background?

Based on this information it may be possible to reduce the transmission quality of objects
outside the area of interest without affecting the viewer’s overall perception of the quality of the
video.

In order to answer these questions, three sets of experiments were carried out.

- **Experiment 1.** The first set of experiments was carried out to identify the regions of high
  and low interest in an e-learning video.
- **Experiment 2.** The second set of experiments was carried out to understand how the
  subjects respond to the change in the perceptual quality of the foreground and
  background.
- **Experiment 3.** The third set of experiments was carried out to understand if the subjects
  notice any addition or removal of the objects in the background.

### 4.3 EXPERIMENTAL PROCEDURE

#### 4.3.a Outline of the Experimental Procedure

In all the three experiments the test sequences were displayed at 25 fps so as to present a
smooth movement. The CIF (352x288 pixels) resolution was maintained because that is a
commonly used size for many commercially available video-conferencing softwares. Each
video sequence was 15 seconds long. The audio was kept muted throughout the sequence to
prevent any interaction between the video and audio.

At the start of the experiments the subjects were asked to imagine that they were students
attending an e-learning class and asked to view the video as they would normally do under such
a situation. They were asked to “pay attention” to the instructor even though they did not really
follow what was being “taught”. Even though this technique has its limitations because there is
a strong interaction between the perception, attention, audio stimulus, familiarity of the subject being taught, etc., this was considered to be the best within the experimental limitations.

At the end of each sequence the subjects were asked to fill in a detailed questionnaire. The questionnaire asked them to describe the scene and then answer a few specific questions. The specific questions covered the details about the room, instructor, objects, colour, location, etc. They were specifically asked not to guess, but rather state ‘don’t remember’ when they had failed to notice some details.

4.3. b Participants

Fifty individuals (nineteen males, thirty one females) in the age group 18-45 years participated in this experiment. All had normal or corrected-to-normal vision. All participants were fully comfortable with the location and adapted to the prevailing illumination conditions before beginning their task. All participants took part in all three video sequences and the order of the experiment was randomised.

4.3. c Viewing Conditions

The video sequences were viewed on a desktop cathode ray tube (CRT) monitor. This monitor was chosen because, in spite of the growing popularity of the Liquid Crystal Display (LCD) monitors, CRT monitors are still relatively cheaper and hence more common for home use. The viewing conditions that were maintained were:

- Display: Cathode ray tube (CRT) desktop monitor
- Size: 36.5 cm x 27.5 cm (46 cm diagonal)
- Resolution: 1024 × 768
- Colour: Quality: Highest (32 bit)
- Distance from eyes: 65 cm (approximately)
- Viewing angles: 90° from vertical

At the start of the experiments the subjects were asked to imagine that they were students attending an e-learning class and asked to view the video as they would normally do under such a situation. They were asked to “pay attention” to the instructor even though they did not really follow what was being “taught”. Even though this technique has its limitations because there is a strong interaction between the perception, attentions, audio stimulus, familiarity of the subject being taught, etc. this was considered the best within the experimental limitations.
4.4 EXPERIMENT 1 – IDENTIFICATION OF THE OBJECTS OF CENTRAL AND MARGINAL INTERESTS

Work done by Rensink et al. [1997] has shown that a picture can be divided into the different regions called the regions of “Central Interest” and “Marginal Interest”. These terms were arbitrarily defined by Rensink et al. [1997] in the context of “Change Blindness” wherein five naive observers were shown an original picture and modified picture interspaced by a gray blank screen. Afterwards the subjects were asked to provide a verbal description of the pictures. From the answers given:

- **Central Interests (CIs)** were defined as objects or areas mentioned by three or more observers;
- **Marginal Interests (MIs)** were objects or areas not mentioned by anyone.

Their work has further shown that subjects tend to notice and remember more details about the features of objects of Central Interest than about the objects of Marginal Interest. It has been proposed that this is because the viewers tend to pay more attention to the objects of Central Interest as compared to the Objects of Marginal Interest.

The informal definition of a foreground is the object at the “front”, “prominent position” or “conspicuous position”; since these descriptions are intangible in this research work it has been proposed that the foreground may be considered as the object of Central Interest and background as the object of Marginal Interest so as to get a quantifiable definition.

In the context of an e-learning video it then becomes essential to scientifically identify foreground (object of Central Interest) and background (object of Marginal Interest).

4.4a Hypothesis

Based on the work of Rensink et al. [1997] on Change Blindness, and Cater et al. [2001, 2002, 2003a, 2003b] on rendering, it was predicted that there were certain regions of Central Interest and some regions of Marginal Interest in a scene. The hypothesis was that the viewers would not pay attention to the regions of Marginal Interest and hence would not remember its attributes.

4.4b Video Sequences

Three video sequences: Head-and-Shoulder, Middle and Panorama were used for the experiments. The three video sequences were of MPEG2 format as downloaded from the camcorder. No additional processing was done on them, so as to minimise the effect of artifacts.
### TABLE 4.1. Results of the experiments carried out to identify the regions of Central Interest and regions of Marginal Interest in the three e-learning video sequences.

<table>
<thead>
<tr>
<th>Question</th>
<th>Number of Correct Answers (Number of participants = 50)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence 1 – Head-and-Shoulder</strong></td>
<td></td>
</tr>
<tr>
<td>What was the colour of instructor’s shirt?</td>
<td>14</td>
</tr>
<tr>
<td>What was the colour of instructor’s collar?</td>
<td>5</td>
</tr>
<tr>
<td>What did instructor have around her neck?</td>
<td>0</td>
</tr>
<tr>
<td>Was instructor wearing spectacles?</td>
<td>32</td>
</tr>
<tr>
<td>Was instructor wearing ear rings?</td>
<td>0</td>
</tr>
<tr>
<td>How was instructor’s hair tied?</td>
<td>0</td>
</tr>
<tr>
<td><strong>Sequence 2 - Middle</strong></td>
<td></td>
</tr>
<tr>
<td>What was the colour of instructor’s dress?</td>
<td>23</td>
</tr>
<tr>
<td>What was the colour of instructor’s collar?</td>
<td>3</td>
</tr>
<tr>
<td>What did instructor have in her hand?</td>
<td>50</td>
</tr>
<tr>
<td>What did instructor draw on the whiteboard?</td>
<td>17</td>
</tr>
<tr>
<td><strong>Sequence 3 - Panorama</strong></td>
<td></td>
</tr>
<tr>
<td>How many whiteboards were there?</td>
<td>50</td>
</tr>
<tr>
<td>What was written on the whiteboard?</td>
<td>36</td>
</tr>
<tr>
<td>Was there a piece of paper on the table?</td>
<td>13</td>
</tr>
<tr>
<td>What was on the side of the table?</td>
<td>0</td>
</tr>
<tr>
<td>What was above the whiteboard?</td>
<td>0</td>
</tr>
<tr>
<td>What was beside the whiteboard?</td>
<td>0</td>
</tr>
</tbody>
</table>

### 4.4. Results

Table 4.1 shows the summary of the results. As can be seen in the Head-and-Shoulder sequence, the only object of Central Interest was the face. Subjects did not remember much about any other object. In the bottom-up processing, subjects pay attention either to the objects that stand out or to the objects that change. In the human face the expression comes from the lips and the eyes hence, eyes tend to focus more on them. As a result, the face attracts maximum attention of the subjects. Hence, the face is the object of Critical Interest and all other objects are of Marginal Interest (Cater. et. al).

In the Middle sequence, all the subjects remembered the instructor going to the whiteboard and writing something on it using a marker. None of them remembered what was written on the board, which was understandable since the writing was totally out of context, but seventeen subjects remembered that she drew an arrow. This could be because that was the only part they understood. In this sequence the subjects overall followed the movements of the instructor and looked at the whiteboard.
In the Panorama sequence, all the subjects remembered was the instructor pacing in front of the whiteboard. Even though they remembered the whiteboard, only thirty six of them could tell that nothing was on the board. Only thirteen subjects could remember a piece of paper on the table. This shows that the degree of attention varies as

\[
\text{Instructor} > \text{Whiteboard} > \text{Paper on the table}
\]

All other objects did not draw any attention. The results may be explained as the instructor and the boards were the objects of Central Interest, even though the writing on the board itself was not. This may be speculated that the subjects could have looked at the board and found nothing on it and hence lost interest in the writing. So, later on they remembered the board without remembering if anything was written on it or not. Thirteen subjects remembered seeing the paper, probably because the white paper stands out on the dark table.

From the results of the experiment carried out, the objects of Central and Marginal Interests were identified and shown in Figure 4.1. Basically the instructor and the whiteboard are the objects of Central Interest. The whiteboard in the Figure 4.1.c, encircled in green, is itself the object of Central Interest, but the writing itself depends upon the interest of the subjects [Werner and Thies 2000].

The results of the experiments confirmed the hypothesis that in an e-learning video sequence there are regions of Central Interest where the viewers pay relatively more attention as compared to the regions of Marginal Interest where they pay less attention. This is reflected by the fact that the viewers remember more about the regions of Central Interest compared to the regions of Marginal Interest.

According to the Merriam-Webster Online Dictionary, *Foreground* is an object in “a position of prominence”, where a prominent object is one that is “readily noticeable”, hence, it can be argued that a “foreground” object is one that is “readily noticeable”. The preliminary results from this set of experiments indicate that the objects of Critical Interest are “readily noticeable”, hence, it may be reasonable to define “foreground” as the objects of Critical Interest.
FIGURE 4.1. Regions of Central and Marginal Interests. The region encircled with red or green show the region of Central Interest.

According to the American Heritage Dictionary, *Background* is "the part of a pictorial representation that appears to be in the distance and that provides relief for the principal objects in the foreground". It has been discussed earlier that the *foreground* can be defined as the object of Critical Interest, hence the "background" can be defined as all the objects that are not of Critical Interest. Hence, background objects are the objects of Marginal Interest.

**Hence, foreground objects are the objects of Critical Interest and the background objects are the objects of Marginal Interest.**
4.5 EXPERIMENT 2 – RELATIONSHIP BETWEEN THE INATTENTIONAL BLINDNESS AND THE PERCEPTUAL QUALITY OF VIDEO

It is well-known to psychologists that humans focus their attention on the objects of interest at the expense of other details; as a result they suffer from Inattentional Blindness. Cater et al. [2002] have shown that in a rendered sequence it is possible to selectively render objects not related to the task at lower resolution, without the viewer noticing any reduction in quality. They have exploited this finding to generate high perceptual quality animated sequences, wherein the non-critical objects were rendered at a lower quality, without having any significant effect on the viewer's perception.

In the previous section, Section 4.4, it was shown that in an e-learning video there exists regions of Central Interest, which may be considered to be the foreground, and regions of Marginal Interest, which may be considered to be the background.

If it could be demonstrated that while watching an e-learning video viewers tend to overlook the degradation in the quality of background as long as the quality of foreground is maintained, then it may be possible to reduce the bandwidth required for streaming an e-learning video. This may be achieved by separating foreground and background and selectively transmitting the foreground at high resolution and the background at low resolution.

4.5.a Hypothesis

The hypothesis was that it would be possible to directly exploit the top-down visual attention process to attract viewers' attention towards the task at hand so that they become inattentive to the quality of the non-critical sections of the video frame. In the previous set of experiments it was observed that in an e-learning sequence the instructor and the whiteboards were the objects of Central Interest and all other portions were of Marginal Interest. Hence, as long as the image of instructor and the whiteboards are maintained at high quality, any degradation in the quality of the rest of the frame would go unnoticed.

4.5.b Video Sequences

The video sequences used for the experiments were prepared by superimposing foreground over background.
The background, the Panorama view of the class without the instructor, was prepared in two qualities - high and low.

- **High Quality Background.** The high quality was a 352x288 pixels still photograph of the Panorama shot without the instructor.

- **Low Quality Background.** The low quality background was generated by first resizing the 352x288 pixels High Quality Background image to 50% of the original (176x144 pixels) and in the second step up-sizing it back to a 352x288 pixels image using Microsoft Photo Editor. This formatting of image in two steps – first reduction in size and then increase in size back to the original size – led to an overall degradation of the quality.

In order to extract the foreground from the video sequence, the High Quality background image was subtracted from the Panorama sequence using the procedure described later in the Sections 4.10 and 4.12. This subtraction identified Foreground as all the objects in the Panorama sequence that were not present in the background. This effectively meant that only the instructor was identified as the foreground. This subtraction gave a video sequence only of the instructor against a blank background.

- **High Quality Foreground.** The high quality foreground was extracted by subtracting High Quality Background from the Panorama video sequence with instructor. It was a sequence of 352x288 pixels frames.

- **Low Quality Foreground.** Low quality foreground was prepared by downsizing the extracted foreground to 50% of the original size (176x144 pixels) and then upsizing it back to the original size (352x288 pixels).

In the next step the sequence containing only the instructor was superimposed over appropriate background to produce the test sequences. Using the two background images and two foreground sequences, four types of video sequences were prepared (**Figure 4.2**).

1. **High Quality Sequence.** High Quality Foreground over High Quality Background.
2. **Low Quality Sequence.** Low Quality Foreground over Low Quality Background.
3. **Composite 1 Quality Sequence.** High Quality Foreground over Low Quality Background.
4. **Composite 2 Quality Sequence.** Low Quality Foreground over High Quality Background.
FIGURE 4.2. The sequences made for studying Inattentional Blindness in e-learning videos. (a) Still from High Quality sequence (b) Still from the Low Quality sequence (c) Still from the Composite 1 Quality sequence and (d) Still from the Composite 2 Quality sequence.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Foreground</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Quality</td>
<td>High (352\times288 \text{ pixels})</td>
<td>High (352\times288 \text{ pixels})</td>
</tr>
<tr>
<td>Low Quality</td>
<td>Low (176\times144 \text{ pixels})</td>
<td>Low (176\times144 \text{ pixels})</td>
</tr>
<tr>
<td>Composite 1 Quality</td>
<td>High (352\times288 \text{ pixels})</td>
<td>Low (176\times144 \text{ pixels})</td>
</tr>
<tr>
<td>Composite 2 Quality</td>
<td>Low (176\times144 \text{ pixels})</td>
<td>High (352\times288 \text{ pixels})</td>
</tr>
</tbody>
</table>
Each test sequence was made by joining two sequences separated by a Head-and-Shoulder sequence of 10 s (Figure 4.3). These experiments are based on a similar work by Cater et al. [2002] and Rensink et al. [1997]. While Rensink et al. [1997] used blank gray screen to separate the two pictures, here a different sequence was used for the separation because that was thought to be closer to the actual e-learning video.

At the beginning of the experiment the subjects were shown the High Quality Sequence and the Low Quality Sequence so as to provide them with the reference sequences. The paired sequences were then shown in a random order.

4.5.c Results

Table 4.2 and Figure 4.4 show the results of the experiment. As is to be expected the participants did not report any difference in the video quality of the High Quality – High Quality and Low Quality – Low Quality sequences because there was none.

When the High Quality – Low Quality and the Low Quality – High Quality sequences were shown all the participants were able to notice the difference. These experiments clearly show that the two video sequences – High Quality and the Low Quality – were significantly different from each other so that all the viewers could notice the differences.
TABLE 4.2. Results of the experiments carried out to study Inattentional Blindness in e-learning video.

<table>
<thead>
<tr>
<th>Test Number</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Same</th>
<th>Better</th>
<th>Worse</th>
<th>Don't Remember</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High</td>
<td>High</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Low</td>
<td>Low</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>High</td>
<td>Low</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Low</td>
<td>High</td>
<td>0</td>
<td>49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Low</td>
<td>Composite 1</td>
<td>8</td>
<td>41</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Composite 1</td>
<td>Low</td>
<td>6</td>
<td>0</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>High</td>
<td>Composite 1</td>
<td>36</td>
<td>0</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Composite 1</td>
<td>High</td>
<td>37</td>
<td>9</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>Low</td>
<td>Composite 2</td>
<td>34</td>
<td>7</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>Composite 2</td>
<td>Low</td>
<td>31</td>
<td>1</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>High</td>
<td>Composite 2</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>Composite 2</td>
<td>High</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

When the Low Quality – Composite 1 Quality and the Composite 1 Quality – Low Quality sequences were shown, most of the people were able to notice the difference. On the other hand, when the High Quality – Composite 1 Quality and the Composite 1 Quality – High Quality sequences were shown, relatively few, 36% and 37% of the viewers, were able to notice the difference.

When the Low Quality – Composite 2 Quality and the Composite 2 Quality – Low Quality sequences were shown, most of the people thought that they were the same. On the other hand, when the High Quality – Composite 2 Quality and the Composite 2 Quality – High Quality sequences were shown, most of the viewers thought the difference was quite significant and they found the Composite 2 sequence rather poor.

These results are interesting because when the quality of foreground changed while the quality of background remained constant, then most of the people were able to notice the difference, but when the quality of background changed while the quality of foreground remained constant then not many people were able to notice the difference.

**Hence, it has been shown that in an e-learning video sequence, as long as the quality of the foreground is maintained at a high level, some degradation in the quality of background does not significantly lower the perceptual quality of the sequence.**

107
FIGURE 4.4. Results of the experiment showing the percentage of viewers who noticed the difference between the qualities of two sequences.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Foreground</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>HQ</td>
<td>High Quality</td>
<td>High (352x288 pixels)</td>
</tr>
<tr>
<td>LQ</td>
<td>Low Quality</td>
<td>Low (176x144 pixels)</td>
</tr>
<tr>
<td>C1Q</td>
<td>Composite 1 Quality</td>
<td>High (352x288 pixels)</td>
</tr>
<tr>
<td>C2Q</td>
<td>Composite 2 Quality</td>
<td>Low (176x144 pixels)</td>
</tr>
</tbody>
</table>

4.6 EXPERIMENT 3 – RELATIONSHIP BETWEEN THE CHANGE BLINDNESS AND THE PERCEPTUAL QUALITY OF VIDEO

Extensive work has been done by Rensink et al. [1997, 2000a, 2000b, 2002] and Simons et al. [1997, 2000, 2005] to understand the conditions under which humans suffer from Change Blindness. They have demonstrated that the intensity of Change Blindness is more for the objects of Marginal Interest as compared to the objects of Central Interest.

In Section 4.2 the definition of “foreground” and “background” in the context of e-learning videos was developed. In the next section, Section 4.3, it was demonstrated that the viewers tend to overlook the degradation in the quality of background as long as the quality of foreground is maintained.
It is envisaged that there may be times when the objects of Marginal Interest may change their position or other attributes. For example, in the classroom setting shown in Figure 3.3, it is worth knowing whether slight changes to the background, like moving of the seconds hand of the wall clock and falling-off of the envelope would be noticed by the viewers. If such changes go unnoticed then it may be possible to save on bandwidth by not updating the background even when there is some change in the background, example, after the envelope has fallen off, or by infrequently updating in the case of regular motion of a non-critical object, example movement of the seconds hand of the wall clock.

4.6.a Hypothesis

The hypothesis was that humans suffer from Change Blindness as a result some change in the objects of Marginal Interest would go unnoticed by the viewers.

4.6.b Video Sequences

The video sequences used for the experiments were prepared by superimposing foreground over background.

The background, the Panorama view of the class without the instructor, was prepared in two qualities – high and low.

- **Original Background.** The high quality was a 352x288 pixels still photograph of the Panorama shot without the instructor (Figure 4.5a).

- **Modified Background.** The modified background was generated by carefully removing some objects from the Original Background using Adobe Photoshop 7.0. The modified background is shown in Figure 4.5b.

In order to extract the foreground from the video sequence, the High Quality background image was subtracted from the Panorama sequence using the procedure described later in the Sections 4.9 and 4.11. This subtraction identified Foreground as all the objects in the Panorama sequence that were not present in the background. This effectively meant that only the instructor was identified as the foreground. This subtraction gave a video sequence only of the instructor against a blank background.

In the next step the sequence containing only the instructor was superimposed over appropriate background to produce the test sequences. Using the two background images and one foreground sequence, two types of video sequences were prepared (Figure 4.5c and 4.5d).
FIGURE 4.5. The sequences made for studying Change Blindness in e-learning videos. (a) Background used for the Reference sequence (b) Background used for the Modified sequence (c) Still from the Reference sequence (d) Still from the Modified sequence.

- **Reference Sequence.** This was made by superimposing foreground over the Original Background.

- **Modified Sequence.** This was made by superimposing foreground over the Modified Background.

Each test sequence was made by joining two sequences separated by a Head-and-Shoulder sequence of 10 s (Figure 4.6). These experiments are based on a similar work by Cater et al. [2000] and Rensink et al. [1997]. While Rensink et al. [1997] used a blank gray screen to separate the two pictures, here a different sequence was used for the separation because that was thought to be closer to the actual e-learning video.
The paired sequences were then shown to the viewers and at the end of the sequence the subjects were asked to tell the differences between the two sequences. The participants were shown the same paired sequences until they noticed the first difference, whichever it was.

4.6. c Results

Table 4.3 and Figure 4.7 show the results of the experiment. The results are interesting because no one was able to notice the difference for the first two times. Thereafter more and more people were able to catch the differences. Out of the fifty, thirty four participants were able to notice the addition of Writing on the Whiteboard as the first difference while the remaining sixteen saw the disappearance of the Paper on the Table as the first difference.

The results of this set of experiments confirmed the hypothesis that when two similar e-learning videos, wherein differences in the objects of Marginal Interest are interspaced by a different video and shown to the viewers, the viewers find it difficult to notice the differences. In other words, humans suffer from Change Blindness as a result they are unable to notice changes in the objects of Marginal Interest.
FIGURE 4.7. Results of the experiment showing the percentage of viewers who notice the difference between the two sequences.

TABLE 4.3. Results of the experiments carried out to study Change Blindness in an e-learning video.

<table>
<thead>
<tr>
<th>Number of times the sequence had to be shown</th>
<th>Number of viewers who spotted the first difference</th>
<th>Percentage of viewers who spotted the first difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
4.7 DISCUSSION ON THE RESULTS OF PSYCHOPHYSICAL EXPERIMENTS

From the psychological experiments carried out, a number of interesting and important observations can be made. The experiments have shown that:

The experiments have provided a scientific basis for categorising “foreground” and “background”. The experiments have shown that “foreground” and “background” may be roughly equated to the “Objects of Central Interest” and “Objects of Marginal Interest”. The objects of Central Interest and Marginal Interest are specific to the person, time, occasion etc. In an e-learning video it has been demonstrated that for the sequences tested, the objects of Central Interest are the instructor and the whiteboard.

It has also been observed that the viewers pay relatively more attention to these two objects and ignore others. As a result, they are often unable to notice any degradation of background as long as the foreground objects are maintained at high quality. It has also been observed that any changes to the background objects go largely undetected.

These observations regarding the behaviour of viewers towards foreground and background offer exiting possibilities for manipulating the two individually so as to obtain the best perceptual quality within the limitation of the available bandwidth.

From the results of experiments carried out to study the role of Inattentional Blindness in the perception of e-learning videos, it may be extrapolated that it might be possible to transmit an e-learning video in which the foreground is transmitted at high resolution and the background is transmitted separately at a relatively lower resolution without compromising the perceived quality at the receiver end.

The results of the experiments on Change Blindness suggest the possibility of transmitting an e-learning video in which the foreground is transmitted at high resolution and the background is transmitted intermittently. Slight anomalies which may arise due to infrequent transmission may go unnoticed by the viewers.
4.8 SEPARATION OF FOREGROUND

Having researched the psychological basis for the separation of foreground and background, the next step is to use the knowledge gained to find the best separation technique. Most of the commonly used foreground separation techniques detect either the relative motion between foreground and background or subtract pre-recorded background frames from the new frame. Hence, it is easy to detect the instructor as the foreground, but other important objects like the whiteboard would ordinarily go undetected as a background. Another problem with most of the commonly used techniques is that the algorithm is normally based on only one attribute – relative motion, or difference, but in the case of e-learning it has been shown that different objects having different attributes may all be identified as foreground objects. In such a scenario it is imperative that two or more relevant techniques be combined in order to achieve the desired result. Hence, a combination of separation techniques, along with human judgment, is required in order to find the complete system. In this research a combination of two different separation techniques has been investigated for the video sequences.

The psychological experiments have shown that there are two foreground objects in the video sequences. These are (a) the instructor; and (b) the whiteboards. The separation of the instructor from a frame can take advantage of the fact that it is possible to take a reference photograph background without instructor and then subtract this reference background from the frames of the video sequences. In order to do this, the suitability of two different algorithms has been studied.

The separation of the whiteboards from the frames exploits the fact that the position of whiteboards does not change with time, hence, if the region occupied by the whiteboards can be identified, it is possible to subtract this fixed region from all the frames of a sequence. This method is based on first identifying the portions covered by various objects in a known background using edge detection techniques, and then using personal judgment to identify objects of Central Interest. The personal judgment is guided by the results of the psychological experiments on the human perceptual system that has identified the whiteboard as the stationary foreground.

At the beginning of the session, these background images will be stored both at the sender and receiver ends. During the lecture, the video will be grabbed and the foreground separated from the frames. The foregrounds so generated will be transmitted to the receivers. At the receiver end the foregrounds will be merged with appropriate backgrounds and shown to the receiver. The quality and the frequency of background transmission can be adjusted so as to
optimise between the degradation of quality and saving of bandwidth. The limit can be set where the perceptual quality is not significantly degraded, but there is a significant saving on bandwidth.

4.9 CHARACTERISATION OF BACKGROUND

In order to efficiently remove the background from a frame by the subtraction technique, the background needs to be properly characterised. This is because due to lighting, camera operation and random variations, the intensity of a particular pixel varies dynamically. For example, Figure 4.8 demonstrates the variation of pixel intensity for six consecutive frames taken from a sequence captured at 25 fps. Figure 4.10a is the frame at a particular instance of time and the Figures 4.8b-f have been obtained by subtracting the next consecutive frames from the frame in Figure 4.10a. In order to make the difference noticeable the images were turned into “negative” and enhanced using the “AutoBalance” feature of the Microsoft Photo Editor Version 3.0.2.3. The figure clearly shows that the random variations in the frames led to small yet significant differences at each pixel. Hence, the statistics of the background need to be studied.

4.9.a Experimental Procedure

A video sequence of 20 seconds of the background only was captured using a camcorder at a rate of 25 frames per second. This sequence was broken into individual PPM frames in order to facilitate easy reading of the RGB intensities of each pixel. Ten locations on the frame were randomly selected and its pixel RGB intensities read.

4.9.b Statistical Analysis

Having obtained the experimental values of the pixel intensities, the statistical analysis was carried out. The mean and standard deviation of pixel RGB intensities of all the 500 frames were calculated.

\[
bm_i = \frac{\sum_{k=1}^{N} b_{k,i}}{N} \quad (4.1)
\]

\[
bxdi = \sqrt{\frac{\sum_{k=1}^{N} (b_{k,i} - bm_i)^2}{N - 1}} \quad (4.2)
\]

\[
bd_{k,i} = b_{k,i} - bm_i \quad (4.3)
\]
FIGURE 4.8. Variations in the intensities of six consecutive frames. (a) The first frame taken as the reference frames. (b-f) Difference between the five frames and the reference frame. The images have been turned into negative and enhanced using the “AutoBalance” feature of Microsoft Photo Editor.
Where

\( b_{ni} = \) Mean value of a pixel of a background. It is calculated for each of red, blue and green intensities.

\( b_{ki} = \) \( i^{th} \) pixel in \( k^{th} \) background frame

\( N = \) Number of frames (here 500)

\( b_{sdi} = \) Standard deviation of a pixel of a background. It is calculated for each of red, blue and green intensities.

\( b_{dk,i} = \) Difference between the intensity value of \( i^{th} \) pixel in \( k^{th} \) background frame and mean value. It is calculated for each of red, blue and green intensities.

**Figure 4.9** is an example of the variation in the intensity values of \( b_{dk,i} \) for a pixel. The figure also shows that the distribution peaks at the mean fall equally on either side. This suggests the possibility of modeling the distribution as a standard Gaussian function.

\[
\text{Probability of a Pixel Being Background Pixel} = \frac{e^{\frac{-{(b_{dk,i} - b_{ni})}^2}{2b_{sdi}^2}}}{b_{sdi}\sqrt{2\pi}} \quad (4.4)
\]

**Figure 4.9** also shows that the pixel values roughly follow the standard Gaussian function. The figure shows that for the particular pixel the intensity varies roughly between \( \pm 5 \) for green and \( \pm 6 \) for blue pixels. The distribution for red is rather skewed and varies between -4 to +11. The distribution of red could have been better modeled using the Weibull distribution function. While Weibull distribution would have improved the modeling results, it would have also increased the computational complexity.
4.10 SEPARATION OF INSTRUCTOR FROM THE FRAMES

The separation of instructor from the frames is based on the premise that it is possible to take the images of the classroom before the instructor walks in. As discussed in the previous section, these background images are then characterised pixel by pixel to calculate the statistics of the background images to account for the random variations. Once the statistics of the background variation have been calculated and the values modelled as the standard Gaussian curve, the thresholds are set, based on the personal judgement. These threshold values define the pixels that can be considered as the background. Figure 4.10 schematically shows this decision-
human visual system based foreground-background separation

4.10. A Figure 4.10. Schematic representation of the decision making process for setting the threshold.

making process. Once the background has been defined, there is a need to separate the foreground from the frames by comparing the frames and the background. This comparison can be done in a number of ways. In this research work, the suitability of two different algorithms for this comparison has been studied.

4.10.a Algorithm X

In Algorithm X, the intensities of every pixel are compared against the intensities of every pixel of a mean background. In case the difference between the two is less than the set threshold, the pixel is considered as the background or else it is considered as the foreground. Figure 4.11 shows the effect of threshold on background separation for different threshold values. The figure shows that with the increase in the threshold value the area identified as the background increases, hence, by using the proper value of threshold it may be possible to separate foreground from a frame.

Figure 4.12 shows the results of the similar calculation for a different frame (Frame 100). The figure shows that, for the same threshold values, the area identified as the background differs considerably from those shown in Figure 4.11. A comparison between Figures 4.11 and 4.12 shows that this simple algorithm cannot be relied upon to correctly identify and separate the background.
FIGURE 4.11. Effect of threshold on the background removal using Algorithm X for Frame 1.
4.10. b Algorithm Y

One of the variations of the Algorithm X has been proposed by Pan et al. [2002]. In the simplified form, the fourth order variance, i.e., fourth power of the difference between the pixel value of the current frame and the mean of the background frames, is calculated. This is calculated according to the formula:

$$\sigma^4 = \frac{\sum (b_{k,i} - bm_i)^4}{N} \quad (4.5)$$

If the value is greater than a set threshold, then the pixel is considered as foreground; otherwise it is taken as background.

Figure 4.13 shows the effect of threshold on background separation. It shows that if the threshold is kept as 250000 then the probability of removing the background is very high.

Like the Algorithm X, this Algorithm also suffers from the same weakness. For example Figure 4.14 shows the effect of threshold on the Frame 200 of the Panorama sequence. The comparison between Figures 4.13 and 4.14 shows that the same threshold values do not work for different frames.

Figures 4.11-4.14 show that the two algorithms remove the background in different ways, and individually they are unable to remove all the background pixels. A better result is obtained by a combination of these two algorithms. The combination has been done using OR operator. If a pixel has been identified as a background using either of the two algorithms then it is a background pixel. Figure 4.15 shows the result of the combination of the two algorithms. In this combination the background is removed by adding a clause that if either of the algorithms identifies a pixel as a background then it is a background.

Figure 4.16 shows that using the two algorithms for the separation of the background it is possible to remove over 80% of the portions of the frame. Hence, if only the instructor is considered as the foreground then a substantial saving of the bandwidth can be achieved by removing the background before transmission.

While the performance of the combination is better than the individual one for the two algorithms, it is still far from perfect. Hence, better algorithms are required for this task.
Threshold = 10000

Threshold = 100000

Threshold = 175000

Threshold = 250000

Threshold = 500000

Threshold = 1000000

**FIGURE 4.13.** Effect of threshold on the background removal using Algorithm Y for Frame 1.
FIGURE 4.15. Effect of threshold on the background removal using a combination of Algorithms X (Threshold = 25) and Algorithm Y (Threshold = 250,000) for Frame 100.
4.10. c Limitations of Algorithms X and Y

The two algorithms suffer from a number of shortcomings. Some of the problems are:

- **Illumination changes.** These can be gradual or sudden. For example, in a classroom illuminated by sunlight there will be a gradual change in the light conditions of the classroom. In an experiment conducted it was found that the pixel values of background changed by approximately 33 counts over a period of 15 minutes.

- **Non-ideal distribution.** Figure 4.9 shows that the experimentally found distribution does not follow the ideal Gaussian distribution, hence, while most of the background will be removed, not all the background pixels will be removed.

- **Wrong interpretations.** If sections of the foreground have values close to that of the background, it will be wrongly interpreted as the background and removed. For example, parts of the hands in the foreground have been wrongly subtracted.

- **Problems due to the Non-ideal Behaviour Camera.** Video taken using an ordinary camera would have some problems that arise due to non-ideal behaviour of the camera operation. Figure 4.16 shows the percentage of portion identified as the background, the white region
in the Figures 4.11-4.14, obtained using a combination of Algorithm X (Threshold = 25) and Algorithm Y (Threshold = 250000), for all the frames of the Panorama sequence. The figure shows a cyclic variation in the percentage of area identified as the background. Most probably this is due to the camera operation.

Various modifications of the simple Gaussian subtraction technique have been proposed. For example, in order to account for the gradual changes in the light condition, a running average of the background is taken, i.e., the previous background frames are taken as the background for the current frame. Suitable weighted average of the background images may also be used where the recent frames have higher weight. Other more advanced techniques based on Gaussian distribution have also been proposed.

4.11 SEPARATION OF WHITEBOARDS FROM THE FRAMES

The algorithm for the separation of the whiteboards from the frames is based on the fact that the whiteboards are immovable; hence, if the regions occupied by the whiteboards are defined, then this fixed region can be subtracted from all the frames of a sequence.

In order to define the areas occupied by various objects, the initial background image is characterised using edge detection techniques. For this purpose a number of techniques are commonly available. Depending upon the type of background, a suitable technique may be selected. Figure 4.17 shows the results of edge detection using various edge detection filters. Once the suitable edge detection technique is selected using this technique, the region of interest is identified. The region of interest, the whiteboard, is then marked and removed from every frame as a foreground object.

Figure 4.18 shows the result of the combination of various foreground separation techniques. The figure shows that by using the subtraction techniques (Algorithm X and Y) (Figures 4.18c and 4.18d), it is possible to separate the first foreground, instructor (Figure 4.18e), and by using the second technique, defining the fixed region of frame as a foreground (Figures 4.18f and 4.18g), it is possible to separate the second foreground (Figure 4.18h). The combination of these two techniques gives the desired method for separating the foregrounds from the background (Figure 4.18i).
FIGURE 4.17. Edge detection of background using various edge detection filters.
FIGURE 4.18. (a) Background, (b) Background with foreground, (c) Separation of the first foreground using Algorithm X, (d) Separation of the first foreground using Algorithm Y, (e) Separation of the first foreground using the combination of Algorithms X and Y. (Cont.)
Chapter 4

Human Visual System Based Foreground-Background Separation

FIGURE 4.18 (Cont.). (f) Identification of edges using Prewit filter in the background image (a), (g) Identification of the second region of interest, (h) Separation of the second foreground, (i) Separation of the two foregrounds.

4.12 ADDITION OF FOREGROUND OVER BACKGROUND

Once the foreground has been separated from the background, the foreground will be sent to the receivers. Since foreground is only a small part of the total frame, this will result in considerable saving of the bandwidth. At the receiver end, the foreground will be added over the pre-stored background to create a complete image. Figure 4.19 shows the addition of foreground over a background. The figure shows that while it is possible to add the foreground over a background, the technique suffers from some serious problems (Figure 4.20).
FIGURE 4.19. (a) Background, (b) Background with foreground, (c) Separation of foreground using the combination of Algorithms X and Y, (d) Addition of a foreground over a background.

1. Due to the changes in the lighting conditions, the difference between the stored and current backgrounds progressively increases; as a result, the addition of current foreground over the stored background is not seamless (Figure 4.20c). It has been proposed that this problem may be corrected by regularly updating the background image.

2. A slight camera shake can result in very annoying double images (Figure 4.20a). This problem may be solved by firmly mounting the camera on a stand.

3. The whole technique is very computation intensive. It is very difficult to adopt this method in real life using an “ordinary” desktop (discussed later in Section 5.3). For this better computational procedures need to be developed in the future.
4. The separation of the foreground and background and their recombination could not be done in the available MPEG-4 since it did not support this feature. The work could only be done in a simulated way by removal and addition from individual PPM frames.

4.13 SUMMARY AND CONCLUSIONS

This chapter discusses how the properties of the human visual system, in particular Inattentional Blindness and Change Blindness, may be exploited to reduce the amount of information that needs to be transmitted while streaming an e-learning video through the Internet.

It has been shown that in an e-learning video there are regions of Central Interest and Marginal Interests which may be considered respectively to be foreground and background. It has been shown that if the foreground is transmitted at a high quality and the background at a relatively low quality, the viewers at the receiver end will consistently fail to notice degradations in the overall quality of the videos. It has also been shown that the viewers may overlook small inconsistencies that may arise due to the infrequent updation of the background in response to the changes to the background.

Based on the knowledge gained from the psychological experiments a technique for the separation of foreground and background has been proposed. The strengths and weaknesses of this technique has also been discussed.
In this chapter the emphasis has been on the separation of the transmission of the background and foreground in order to reduce the amount of information that needs to be transmitted. In the next chapter the efficient transmission of the data within the available bandwidth shall be discussed.
CHAPTER 5

OPTIMISED TRANSMISSION OF VIDEO IN LAYERS USING MULTI-LAYER FEED-FORWARD CONTROLLER
CHAPTER 5

OPTIMISED TRANSMISSION OF VIDEO IN LAYERS USING MULTI-LAYER FEED-FORWARD CONTROLLER

5.1 INTRODUCTION

In the previous chapter, Chapter 4, a method for the separation of video into foreground and background has been described. This video has to be streamed to the network wherein the receivers are connected by varied bandwidth capacities. It is difficult to satisfy the requirements of a heterogeneous network. One of the methods that have been proposed is the layered transmission of video. For this scheme to be successful and accepted by end-users, it is important that the division and transmission of video is not only based on the requirements of the network but also takes into consideration the psycho-physical behaviour of the human visual system.

This chapter describes the working of the Multi-Layer Feed Forward Controller. The controller dynamically characterizes the nature of the video sequence and combines the knowledge of the bandwidth available to various receivers and then divides the video sequence into layers in an optimal way before transmitting it into the network.

5.2 FUNCTIONAL ARCHITECTURE OF THE MULTI-LAYER FEED-FORWARD CONTROLLER

Figure 5.1 shows the block diagram for the architecture. The investigation was carried out using a commercially available H.261 based video conferencing tool. The tool was modified at the system level to include a filter (a demultiplexer written in Java) at the post codec level. The sender sends the video stream to an Intra-H.261 codec from where the encoded stream goes to a demultiplexer, where the header of the RTP packets are read. On the basis of macroblock address, the demultiplexer partitions the stream and sends them to different multicast addresses.
In the present investigation the video were sent in 1 to 6 layers. Each layer of the coded video was sent on an unique IP multicast address. At the receiver end, the multiplexer combines all the accepted layers and sends them to the decoder.

A stream from encoder is sent to the Multi-Layer Feed Forward Controller to:

1. Dynamically characterise the nature of the video sequence (number of Macroblocks per frame). (Discussed in Section 5.3)
2. Use linear prediction for predicting the complexity of the incoming movie sequence. While the predicted value of the complexity is being used by the controller, the actual values are stored. (Discussed in Section 5.4)
3. Take the input regarding the availability of bandwidth to different receivers.
4. Decide the different possible ways of configuration (number of layers, distribution of frames among different layers and frame rate) of the codec for the incoming video sequence. (Discussed in Section 5.6)
5. Select the best option for the optimum transmission of video. (Discussed in Section 5.6).
6. Send the optimised values of configuration parameters to the demultiplexer.
7. When the encoded stream goes to a demultiplexer, the header of the RTP packets are read. On the basis of frame numbers the demultiplexer partitions the stream according to the recommendations made by the controller and sends them to different multicast addresses.
Chapter 5 Optimised Transmission of Video in Layers Using Multi-Layer Feed-Forward Controller

FIGURE 5.2. Flow diagram showing the basic decision making procedure for the controller.

8. At the receiver end, the multiplexer combines all the accepted layers and sends them to the decoder.

Figure 5.2 schematically shows the flow diagram of the working of the controller.

5.3 DYNAMIC CHARACTERISATION OF THE ONGOING VIDEO SEQUENCE

In this work the complexity of video sequence has been defined as the number of Macroblocks sent per frame. The controller reads the headers of the RTP packets and then counts the number of Macroblocks in a frame. This gives the number of Macroblocks in a frame.
5.4 PREDICTION OF THE QUALITY OF INCOMING VIDEO FRAMES

Figure 5.3 shows schematically the working of complexity calculator. In the complexity calculator the number of Macroblocks per frame is read for a set of frames. At the first instance the number of frames in a set is kept at six. Using the standard least squares fitting technique, the best-fit linear curve for the six points is found. This equation is then used to predict the expected number of Macroblocks per frame for the seventh frame. This predicted value is used as the average complexity for the next set of frames for optimising and finding the best layering technique. While the predicted value is being used by the controller, the next set of actual complexity is stored and used to calculate the subsequent predicted value. The number of frames in the next set is equal to the number of frames being sent in the current layering scheme.

1 The linear least squares fitting technique is a simple and commonly applied form of linear regression and provides a solution to the problem of finding the best-fitting straight line through a set of points.
As an example, Figure 5.4 shows complexity (number of Macroblocks per frame) versus frame number of a video sequence of approximately 2 minutes. The figure also shows the prediction for the video sequence for the fixed number (4) of frames per set. The figure shows that there is a reasonably close match between the actual complexity of the video sequence and the predicted complexity.

5.5 DEVELOPMENT OF THE CONTROLLER EQUATIONS

5.5.a Distribution of Frames Between Layers

As discussed earlier in Section 3.7, in this research the effect of distribution of the frames between the layers using two different algorithms were studied (Figure 3.4).

Figure 5.5 shows the experimental results of the effect of number of layers on the bandwidth requirement. As is to be expected, in the case of Algorithm A the bandwidth requirement increases almost linearly with the increase in the number of layers accepted. In the case of Algorithm B the bandwidth requirement increases in proportion to the $2^{(n-1)}$, where $n$ is the number of layers accepted by a receiver.
5.5. Effect of number of layers on the load on network for a video sequence (oscillating pendulum).

5.5.b Relationship between Complexity - Bitrate - Number of Layers Sent - Number of Layers Received

As has been discussed in the previous section (Section 3.11a), there is a linear relationship between the complexity (number of Macroblocks per frame) and the bandwidth requirement. Hence:

\[
\text{Bitrate} = \text{Const.} \times \text{Complexity}
\]

While developing the equations relating the bandwidth requirement and the number of layers, a simplification has been made. In the Section 3.7 it has been discussed that as the number of layers increases the overhead required for the transmission of video also increases. Here it has been assumed that there is no additional overhead as a result of layering. This has been done in order to keep the equations simple.
For Algorithm A

The bandwidth required for receiving one layer:

\[ BR_1 = \frac{\lambda \times C}{N_{LS}} \quad (5.1) \]

The bandwidth required for receiving \( N_{LR} \) layers:

\[ BR_{LR} = \frac{\lambda \times C \times N_{LR}}{N_{LS}} \quad (5.2) \]

Where:

- \( BR_1 \) = Bitrate required for one layer.
- \( BR_{LR} \) = Bitrate required for accepting \( N_{LR} \) layers.
- \( N_{LS} \) = Number of layers sent
- \( N_{LR} \) = Number of layers received
- \( C \) = Complexity of the video sequence (number of Macroblocks per frame)
- \( \lambda \) = Experimentally found constant

For Algorithm B

The bandwidth required for receiving one layer:

\[ BR_1 = \frac{\lambda \times C}{2^{N_{LS}}} \quad (5.3) \]

The bandwidth required for receiving a number of layers:

\[ BR_{LR} = \frac{\lambda \times C \times 2^{N_{LR}}}{2^{N_{LS}}} \quad (5.4) \]

Where:

- \( BR_1 \) = Bitrate required for one layer.
- \( BR_{LR} \) = Bitrate required for accepting \( N_{LR} \) layers.
- \( N_{LS} \) = Number of layers sent
- \( N_{LR} \) = Number of layers received
- \( C \) = Complexity of the video sequence (number of Macroblocks per frame)
- \( \lambda \) = Experimentally found constant

The values of \((\lambda \times C)\) has to be found experimentally.
When the video sequence is transmitted in one layer:

\[ BR_{LR} = \frac{80 \times N_{LR}}{N_{LS}} \]  

**Algorithm B**

\[ BR_{LR} = \frac{80}{2^{(N_{LR} - N_{LS})}} \]

**FIGURE 5.6.** Examples of results showing the development of the relationship between Bitrate – Number of Layers Sent – Number of Layers Received (Complexity = 4, Frame rate = 3 fps, PQ = Perceptual quality, LN = Load on network, LS = Number of layers sent, LR = Number of layers received)

When the video sequence is transmitted in one layer:

\[ BR_{LR} = \lambda \times C \] \hspace{1cm} (5.5)

Where:

- \( BR_{LR} \) = Bitrate required for transmitting and receiving one layer.
- \( \lambda \times C \) is the bandwidth required for transmitting a video sequence in one layer.

The \( BR_{LR} \) can be measured using any suitable technique. In this research the file size of the video sequence was recorded and divided by the length of the video sequence. As an example, when a video sequence (Complexity = 4) was transmitted at 3 fps, the sequence required 80 kbps bandwidth. Thus, for this sequence and rate of transmission \( \lambda \times C = 80 \) **(Figure 5.6).**
### TABLE 5.1. Effect of complexity and number of layers on the bandwidth requirement.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Frame Rate</th>
<th>Complexity</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>2</td>
<td>( BR_{LR} = \frac{44 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>4</td>
<td>( BR_{LR} = \frac{53 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>6</td>
<td>( BR_{LR} = \frac{72 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>8</td>
<td>( BR_{LR} = \frac{80 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>2</td>
<td>( BR_{LR} = \frac{51 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>4</td>
<td>( BR_{LR} = \frac{103 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>6</td>
<td>( BR_{LR} = \frac{179 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>8</td>
<td>( BR_{LR} = \frac{270 \times N_{LR}}{N_{LS}} )</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>2</td>
<td>( BR_{LR} = \frac{44}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>4</td>
<td>( BR_{LR} = \frac{53}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>6</td>
<td>( BR_{LR} = \frac{72}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>8</td>
<td>( BR_{LR} = \frac{80}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>2</td>
<td>( BR_{LR} = \frac{51}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>4</td>
<td>( BR_{LR} = \frac{103}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>6</td>
<td>( BR_{LR} = \frac{179}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>8</td>
<td>( BR_{LR} = \frac{270}{2^{(N_{LS} - N_{LR})}} )</td>
</tr>
</tbody>
</table>
Figure 5.6 is an example of variation of the bitrate required versus number of layers sent and number of layers received. As is to be expected, for Algorithm A the bitrate increases linearly with the increase in number of layers received and inversely as the number of layers sent and for Algorithm B the bitrate increases exponentially with the increase in the number of layers sent and the number of layers received. Curve fitting was carried out using the Equations 5.2 and 5.4.

Table 5.1 gives the experimental values for two different frame rates. Based on more experiments more values can be found.

5.5. c Relationship between Complexity – Perceptual Quality – Number of Layers Sent – Number of Layers Received

In order to develop the relationship between the perceptual quality of video and the number of layers sent and received, extensive subjective tests were carried out. The perceptual quality of video was subjectively tested according to the recommendations of ITU-T P910 (09/99) standard. This standard was considered to be appropriate in the absence of any other standard made specifically for the video conferencing tools. For each type of video sequence 15 people were asked to give their opinion. The subjects were all postgraduate students from various areas of engineering. They all had over 5 years of experience in computer and Internet operation. Audio was muted in the experiment. A slight modification to the standard was made to make the judgement finer, accordingly the subjects were asked to rate the video sequences in the quality scale between 1-10, instead of 1-5. The subjects were asked to consider the unlayered video displayed at the sender end just before being sent to the network as having quality 10, and all other video sequence were judged in comparison to this sequence.

Figure 5.7 shows a still from a video sequence used for developing the relationship between the perceptual quality of video and the number of layers sent and received. The figure shows the fall in quality due to the formation of blockiness when all the layers of the sequence were not received. With the decrease in the number of fractional layers accepted the blockiness increases due to the fact that even though the blocks had changed the blocks were not replenished for some time.

Figure 5.8 shows the effect of layering on the perceptual quality of video for Algorithms A and B. As is to be expected Algorithm A gives a linear increase in perceptual quality with the increase in number of layers accepted by a receiver. But Algorithm B gives exponential increase in quality with the increase in number of layers accepted by a receiver.
Chapter 5  Optimised Transmission of Video in Layers
Using Multi-Layer Feed-Forward Controller

(a) 6 layers sent 6 layers received
(b) 6 layers sent 1 layer received.

**FIGURE 5.7.** An example of frame from video sequence, showing the effect of number of layers received by the receiver.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Frame rate</th>
<th>Perceptual Quality of Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm A</td>
<td>3 fps</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Algorithm B</td>
<td>3 fps</td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

Algorithm A
Frame rate = 3 fps

Algorithm B
Frame rate = 3 fps

**FIGURE 5.8.** Effect of number of layers sent and number of layers received on the perceptual quality of video (Complexity = 4 Macroblocks/frame).
In order to develop a relationship between the perceptual quality, number of layers sent and number of layers received, the experimental data was converted into equations using curve fitting. For the curve fitting it was assumed that the perceptual quality of video is proportional to the bandwidth. This assumption is reasonably correct for the temporal layering on which the present controller is based. Figure 5.9 shows an example of the effect of layering on the bandwidth allocation and perceptual quality. Figure shows that with the addition of layers, in the case of Algorithm A, the perceptual quality improves uniformly and in the case of Algorithm B the perceptual quality improves exponentially. Hence, it seems reasonable to model the Perceptual quality – Number of layers sent – Number of layers accepted in the same form as the Bandwidth – Number of layers sent – Number of layers.
Chapter 5

Optimised Transmission of Video in Layers Using Multi-Layer Feed-Forward Controller

Figure 5.10 shows the effect of video complexity (number of Macroblocks per frame) on the perceptual quality of the video. It shows that with an increase in the complexity the subjects noted a deterioration in the quality even when the receiver accepted all the fractional layers. In the absence of any detailed study, the reason for this observation may only be guessed. It is possible that the video sequence with low complexity need less bandwidth, hence, they do not suffer much loss of quality during the transmission. Secondly, the slow rate of changes in the sequence having low complexity makes the movement seem smoother to the viewers.
For Algorithm A

Based on Equation 5.6 the bandwidth required for receiving a number of layers:

\[ P_{QLR} = \frac{\alpha \times C \times N_{LR}}{N_{LS}} \]  

(5.6)

Where:

- \( P_{QLR} \) = Perceptual quality achieved by accepting \( N_{LR} \) layers.
- \( N_{LS} \) = Number of layers sent
- \( N_{LR} \) = Number of layers received
- \( C \) = Complexity of the video sequence (number of Macroblocks per frame)
- \( \alpha \) = Experimentally found constant

In order to account for the discrepancies arising due to the effect of complexity on the perceptual quality, the curve fitting was carried out according to the equation:

The bandwidth required for receiving a number of layers:

\[ P_{QLR} = \frac{\alpha \times C \times (N_{LR} + \beta_1)}{(N_{LS} + \beta_2)} \]  

(5.7)

Where:

- \( P_{QLR} \) = Perceptual quality achieved by accepting \( N_{LR} \) layers.
- \( N_{LS} \) = Number of layers sent
- \( N_{LR} \) = Number of layers received
- \( C \) = Complexity of the video sequence (number of Macroblocks per frame)
- \( \alpha, \beta_1, \beta_2 \) = Experimentally found constants

In order to find the values of \( \alpha, \beta_1 \) and \( \beta_2 \) the experimental data correlating Complexity – Perceptual Quality – Number of Layers Sent – Number of Layers Received (as described in Section 5.5.c) is plotted. Using the curve fitting, the best fit curve satisfying the Equation 5.7 is obtained. From this best-fit curve the values of \( \alpha, \beta_1 \) and \( \beta_2 \) are obtained. Figure 5.11 shows examples for the curve fitting found using this model. The figure shows a reasonably good match between the experimental results and the curves.
Algorithm A
\[ PQ_{LR} = \frac{7.8 \times (N_{LR} + 3.1)}{(N_{LS} + 2.9)} \]

Algorithm B
\[ PQ_{LR} = 7.7 \times 2^{(N_{LR} - N_{LS})} \]

**FIGURE 5.11.** Examples of results showing the development of equation between Perceptual Quality – Number of Layers Sent – Number of Layers Received. (Complexity = 4, Frame rate = 3 frames / s)

For Algorithm B

The bandwidth required for receiving a number of layers:

\[ BR_{LR} = \gamma \times C \times 2^{N_{LR}} \frac{2^{N_{LS}}}{2^{N_{LS}}} \]  

(5.8)

Where:

- \( PQ_{LR} \) = Perceptual quality achieved by accepting \( N_{LR} \) layers.
- \( N_{LS} \) = Number of layers sent
- \( N_{LR} \) = Number of layers received
- \( C \) = Complexity of the video sequence (number of Macroblocks per frame)
- \( \gamma \) = Experimentally found constant

In order to find the values of \( \gamma \times C \) the experimental data correlating Complexity – Perceptual Quality – Number of Layers Sent – Number of Layers Received (as described in Section 5.5.c) is plotted. Using the curve fitting, the best fit curve satisfying the Equation 5.8 is obtained. **Figure 5.11** shows examples for the curve fitting found using this model. **Table 5.2** gives the relationship developed for Perceptual video quality – Number of layers sent – Number of layers accepted.
### TABLE 5.2. Effect of complexity and number of layers on the perceptual quality of video.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Frame Rate</th>
<th>Complexity</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>2</td>
<td>$PQ_{LR} = \frac{9.0 \times (N_{LR} + 1.0)}{(N_{LS} + 1.0)}$</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>4</td>
<td>$PQ_{LR} = \frac{7.8 \times (N_{LR} + 1.5)}{(N_{LS} + 1.4)}$</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>6</td>
<td>$PQ_{LR} = \frac{7.7 \times (N_{LR} + 1.2)}{(N_{LS} + 1.2)}$</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>8</td>
<td>$PQ_{LR} = \frac{7.8 \times (N_{LR} + 3.1)}{(N_{LS} + 2.9)}$</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>2</td>
<td>$PQ_{LR} = \frac{6.7 \times (N_{LR} + 0.6)}{(N_{LS} + 0.9)}$</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>4</td>
<td>$PQ_{LR} = \frac{7.2 \times (N_{LR} + 2.6)}{(N_{LS} + 3.0)}$</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>6</td>
<td>$PQ_{LR} = \frac{5.9 \times (N_{LR} + 1.1)}{(N_{LS} + 2.0)}$</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>8</td>
<td>$PQ_{LR} = \frac{8.5 \times (N_{LR} + 3.9)}{(N_{LS} + 3.5)}$</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>2</td>
<td>$PQ_{LR} = \frac{8.8}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>4</td>
<td>$PQ_{LR} = \frac{8.2}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>6</td>
<td>$PQ_{LR} = \frac{8.0}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>8</td>
<td>$PQ_{LR} = \frac{7.7}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>2</td>
<td>$PQ_{LR} = \frac{9.1}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>4</td>
<td>$PQ_{LR} = \frac{8.9}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>6</td>
<td>$PQ_{LR} = \frac{8.7}{2(N_{LS} - N_{LR})}$</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>8</td>
<td>$PQ_{LR} = \frac{7.7}{2(N_{LS} - N_{LR})}$</td>
</tr>
</tbody>
</table>
5.6 CALCULATION FOR THE SELECTION OF THE BEST LAYERING TECHNIQUE

The first step in the decision making process of the controller is finding out the different options (Algorithm to use, Number of layers to be sent and Frame rate) that can be used for transmitting a video sequence to a receiver. From these options the combination that gives the highest perceptual quality is then selected.

In the first step, the controller characterises the ongoing sequence in terms of the complexity (number of Macroblocks per frame) and then predicts the complexity of the incoming sequence. The controller takes in the predicted complexities and bandwidth-available to a receiver as input. Using the relationship between Complexity - Number of layers sent - Number of layers received - Bandwidth it calculates the number of layers the receiver can accept if the transmission is done in 1 to 6 layers. Next using the relationship between Complexity - Number of layers sent - Number of layers received - Perceptual quality the expected perceptual quality can be calculated for the different cases of number of layer sent and corresponding number of layer received.

The above paragraph can be explained in the following points little elaborately:

1. Read the available bandwidth at the receiver end.
2. Use linear prediction for predicting the complexity of the movie sequence.
3. Use the equations developed for the controller (Discussed in Section 5.5.b).
4. For Algorithm A, calculate the maximum number of layers that can be accepted by a receiver if the video is sent in different number (1 to 6) of layers. If any receiver can accept less than 1 layer then that option is not acceptable.
5. For Algorithm A, calculate the expected perceptual qualities for these six options.
6. Repeat the previous steps (Step 5) for Algorithm B.
7. Calculate steps 4, 5 and 6 for different frame rates.
8. Find the option that gives the highest perceptual quality for the receiver.
5.7 AN EXAMPLE OF THE CALCULATION FOR THE SELECTION OF THE BEST LAYERING TECHNIQUE

For example, if the predicted complexity of a video sequence is 8 and if a receiver has available bandwidth of 40 kbps at a particular instance, then the available options are:

Using Algorithm A (For Complexity 8, Frame Rate = 3)

\[ BR_{LR} = \frac{80 \times N_{LR}}{N_{LS}} \]

\[ 40 = \frac{80 \times N_{LR}}{N_{LS}} \]

<table>
<thead>
<tr>
<th>No. of Layers Sent ((N_{LS}))</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum No. of Layers Received ((\text{Max. } N_{LR}))</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
</tr>
<tr>
<td>Actual Maximum (N_{LR}) that the receiver can accept ((\text{mod Max. } N_{LR}))</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Hence, there are six options by which the sender can send the video sequence using Algorithm A at Frame rate = 3 to a receiver connected by 40 kbps connection. For these five options the calculated perceptual qualities are:

Using Algorithm A (For Complexity 8, Frame Rate = 3)

\[ PQ_{LR} = \frac{7.8 \times (N_{LR} + 3.1)}{(N_{LS} + 2.9)} \]

<table>
<thead>
<tr>
<th>No. of Layers Sent ((N_{LS}))</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Maximum (N_{LR}) that the receiver can accept ((\text{mod Max. } N_{LR}))</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Maximum Perceptual Quality</td>
<td>0</td>
<td>6.5</td>
<td>5.4</td>
<td>5.4</td>
<td>5.0</td>
<td>5.3</td>
</tr>
</tbody>
</table>
TABLE 5.3. Calculated options for a receiver having bandwidth of 40kbps for a video sequence whose predicted complexity is 8.

<table>
<thead>
<tr>
<th>Algo.</th>
<th>Frame rate</th>
<th>Equation Used</th>
<th>NLS</th>
<th>NLR</th>
<th>Equation Used</th>
<th>PQLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 3</td>
<td>BR_{LR} = \frac{80 \times N_{LR}}{N_{LS}}</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7.8 \times \left(\frac{N_{LR} + 3.1}{N_{LS} + 2.9}\right)</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td>6.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>2</td>
<td>2</td>
<td></td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2</td>
<td>2</td>
<td></td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>2</td>
<td>2</td>
<td></td>
<td>5.3</td>
</tr>
<tr>
<td>A 8</td>
<td>BR_{LR} = \frac{270 \times N_{LR}}{N_{LS}}</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8.5 \times \left(\frac{N_{LR} + 3.9}{N_{LS} + 3.5}\right)</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>B 3</td>
<td>BR_{LR} = \frac{80}{2^{(N_{LS} - N_{LR})}}</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7.7 \times 2^{(N_{LS} - N_{LR})}</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>3</td>
<td>3</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>5</td>
<td>5</td>
<td></td>
<td>3.8</td>
</tr>
<tr>
<td>B 8</td>
<td>BR_{LR} = \frac{270}{2^{(N_{LS} - N_{LR})}}</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7.7 \times 2^{(N_{LS} - N_{LR})}</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>

Similarly, all the options for the Algorithm A and B for frame rates 3 fps and 8 fps are calculated. Table 5.3 gives the summary of all the available options. In this table it can be seen that the option of Algorithm A, 3 fps, N_{LS} = 2, N_{LR} = 1 gives the highest perceptual quality. Hence, for a receiver connected by 40 kbps bandwidth the best option to receive a video sequence having predicted complexity of 8 is to receive 1 layer when the sender transmits the video sequence in two layers using Algorithm A.

Similar, calculations can be done for all the receivers. The option that will give the highest perceptual quality to a group of receivers is the best option.
5.8 EVALUATION OF THE PERFORMANCE OF THE MULTI-LAYER FEED-FORWARD CONTROLLER

In order to prove the viability of the controller, its performance needs to be evaluated extensively using simulation and experiments in real network. Various parameters that need to be studied are:

- Configuration
- Scalability
- Bandwidth of different links
- Packet loss

Unfortunately, the current network does not support layering, hence, the performance of the controller could not be evaluated in the real network. As a result, the evaluation was only done in a pseudo-real environment in a small laboratory setup.

5.8.1 Experimental Set-up for the Evaluation of the Performance of the Multi-Layer Feed-Forward Controller

The evaluation of the performance of the controller was carried out in a small laboratory set-up. The sender-receiver system comprised of a commercially available videoconferencing tool. The tool used Intra H.261 as the video codec. A post-codec Multi-Layer Feed-Forward Controller as described earlier was attached to it.

The set-up consisted of a sender connected to a router by a 100Mbps link, and the router was in-turn connected to four parallel receivers. The router was a Linux system, in which the bandwidth was regulated using Linux 2.2 Packet Shaping code. The bandwidths four router-receivers links were set as 20, 40, 60 and 80 kbps. At the receiver end required number of layers were merged together to form complete sequence.

Two different types of configurations were tested (Figure 5.12).

- **Configuration 1**: There was one sender and one receiver. Only one receiver was connected to the router and the link capacity was varied in each experiments. Four different link capacities that were tested were 20, 40, 60 and 80 kbps.
- **Configuration 2**: There was only one sender but four receivers. All the receivers were connected in parallel and had link capacities of 20, 40, 60 and 80 kbps.
Chapter 5 Optimised Transmission of Video in Layers
Using Multi-Layer Feed-Forward Controller

For evaluating the performance of the controller, experiments were carried out using three
different combinations of the algorithm:

(a) Using only Algorithm A
(b) Using only Algorithm B
(c) Using both Algorithms A and B

5.8.b Video Sequence Used for the Evaluation of the Performance of the Multi-
Layer Feed-Forward Controller

A video sequence of 2.5 minutes was sent through the set-up and the output recorded at the
receiver end. The movie sequence was so chosen as to have sections having low (2 Macroblocks
per frame) to high (15 Macroblocks per frame) complexity.

5.8.c Subjective Method Used for Evaluating the Quality of the Video Sequences.

The perceptual quality of video was tested subjectively according to the recommendations of
ITU-T P910 (09/99) standard and described in Section 5.5.c.

5.8.d Effect of Bandwidth on the Quality Uncontrolled Video Transmission

When the test sequence was sent to the receiver, through the router, as is to be expected, the
quality of video sequence deteriorated with the fall in the available (80, 60, 40 and 20 kbps)
bandwidth. With the fall in the bandwidth the motion became jerkier and the frames became
blockier. This was due to the random packet drops that took place.

FIGURE 5.12. Configurations used for testing the controller.
Chapter 5 Optimised Transmission of Video in Layers Using Multi-Layer Feed-Forward Controller

5.8. e Evaluation of the Performance of the Controller for Configuration 1

Using Only Algorithm A

For Configuration 1, when the controller was run using only Algorithm A, for the 40 and 60 kbps links, the subjects reported an improvement in the quality of the video sequence with layering compared to video sequence without layering (Figure 5.13). For the 20 and 80 kbps links, the subjects did not report any improvement in the quality for the video sequence with layering compared to the video sequence without layering. The lack of improvement in video quality upon layering in the case of 20 kbs bandwidth link may be because of two reasons. Firstly, the quality of the video sequence in both the cases was so bad that the subjects could not see any difference. Secondly, at high complexity the bandwidth requirement even for one layer may be more than the available 20 kbs resulting in further drop of packets. On the other hand in the case of 80 kbs the quality of both was similar because the available bandwidth was sufficient for the transmission of the video with or without the controller.

Using Only Algorithm B

Figure 5.13 shows the effect of using only Algorithm B on the perceptual quality of video. The subjects reported an improvement in the quality of video sequence with layering compared

![Graph showing improvement in perceptual quality of video](image)

**FIGURE 5.13.** Improvement in the perceptual quality of video in the scale −5 to +5 on the use of controller. Unlayered video has a reference value 0.
Chapter 5

Optimised Transmission of Video in Layers
Using Multi-Layer Feed-Forward Controller

to video sequence without layering for the 20, 40 and 60 kbps links. The improvement in quality at the 20 kbps may be because the Algorithm B works better for the lower bandwidths. At 40 kbps, Algorithm A performed better than Algorithm B, the explanation for this can be seen in the Table 5.5, which discusses a similar example. At 60 kbps both the algorithms perform similarly. At 80 kbps link there were no packet drops due to congestion, hence, there was no effect of layering.

Using Both Algorithms A and B

Figure 5.13 shows the effect of using both Algorithms A and B on the perceptual quality of video. In this case, for the 20, 40 and 60 kbps links, subjects reported better performance for the video sequence with layering compared to without layering. They did not report any improvement in the performance for 80 kbps link. This may be because at the 80 kbps link again there were no packet drops, hence, no effect of layering.

Figure 5.14 is a representative example of the recommendations made by the controller for a portion of the test video sequence. The lower section of the graphs depicted in Figure 5.14a shows the prediction of the complexity of the coming frames. The black line shows the real value of the complexity of the frames, and the red dotted line shows the predicted value. The accuracy of the prediction varies but overall it seems quite close to the original.

Figure 5.14b shows that when the receiver was connected at 20 kbps bandwidth in Configuration 1. The controller used both the Algorithms A and B and the controller selected them automatically. The controller ran the algorithm and decided that the optimal number of layers to be sent was three and the receiver joined Layer 1 and 2 at different times.

Figure 5.14 clearly shows the switching between the two algorithms in the controller based on the complexity at a given instance. Comparison between the two sections of the figure shows that at lower complexity, hence, lower number of Macroblocks per frame, the controller sends the video using Algorithm A. At higher complexity, hence, higher number of Macroblocks per frame, the controller sends the video using Algorithm B.

5.8. Evaluation of the Performance of the Controller for Configuration 2

Both Algorithm A and B

When the controller operates both the Algorithms A and B, then the controller toggles between the two algorithms and the number of layers vary between 3 and 6. Figure 5.15a shows the actual complexity of video sequence and the predicted complexity with respect to the frame number. Figures 5.14.b-5.14.f show the number of layers various receivers join with respect to the frame number.

156
FIGURE 5.14. Output from the controller. The lower portion of the graph shows the actual complexity of the video sequence and the predicted complexity as calculated by the controller. The upper portion of the graph shows the decision made by the controller for Configuration 1 connected with 20 kbps link and Configuration 2. A and B are the algorithms being used at that frame number.
**Figure 5.14f** also shows the variations in the number of layers transmitted at different frame number for the Configuration 2 since all the layers transmitted are also received by the received connected by 80kbps link. The figure shows that the controller adjusts the transmission based on the requirements of all the four receivers. From this optimised transmission different receivers can accept different number of layers. The number of layers the receivers will accept depends upon their bandwidth.

### 5.9 SUMMARY AND CONCLUSIONS

Research was carried out on optimised dynamic layering of video sequence using a post-codec Multi-Layer Feed-Forward Controller. The controller took into account: (a) complexity of the scene (number of Macroblocks per frame) and (b) bandwidth requirement; and optimised (a) number of layers, (b) distribution of frames among different layers, and (c) frame rate so as to provide maximum perceptual quality to the receivers. The controller was evaluated in a pseudo-real network for two different configurations: (a) one sender and one receiver; and (b) one sender and many receivers connected in parallel. It was observed that the quality of transmission improved by the use of the controller.

From the work carried out it may be concluded that it is possible to make a simple post-codec controller that can divide a video sequence into layers before sending to the network. By properly controlling the layering procedure it is possible to improve the quality of video at the receiver end by replacing random packet drops with more controlled packet rejections.

The work was stopped midway because it was realised that practical implementation of any layering scheme required considerable changes in the hardware of the existing network. These changes in the network were not expected to take place in the near future. Hence, at this stage, the concept of Feed-Forward Controller was modified so as to optimise the transmission of video in a single layer in a perceptually optimised way. This concept need not be dismissed altogether because it could be developed in the future.
CHAPTER 6

DESIGN OF SINGLE-LAYER FEED-FORWARD CONTROLLER
CHAPTER 6

DESIGN OF SINGLE-LAYER FEED-FORWARD CONTROLLER

6.1 INTRODUCTION

The main idea of this research work is to improve the perceptual quality of video at the receiver end by efficiently utilising the available bandwidth. The previous chapter (Chapter 4) discussed the possibility of reducing the amount of data that needs to be transmitted, while maintaining the perceptual quality of video. In this chapter, efficient utilisation of the bandwidth shall be discussed.

In this research work it has been proposed that the use of a Single-Layer Feed-Forward Controller to vary the quality of video according to the availability of bandwidth would help in an efficient utilisation of the available bandwidth. This controller dynamically characterises the ongoing video sequence for its Spatial and Temporal Information, then predicts the quality of the incoming sequence; based on those results it adapts the transmission of video in terms of frame-rate and resolution to efficiently utilise the available bandwidth in order to provide a perceptually optimal quality of video at the receiver end.

The suitability of a video transmission system is judged by its friendliness to the network and the satisfaction of the users. This evaluation of the proposed scheme is discussed in the next chapter, Chapter 7.

6.2 FUNCTIONAL ARCHITECTURE OF THE SINGLE-LAYER FEED-FORWARD CONTROLLER

In the proposed architecture (Figure 3.1), the video signals from cameras are taken and encoded using an open source-code MPEG-4 software (FFMPEG) [FFMPEG Multimedia System 2006] on Linux (Fedora Core 2) platform. The encoded video is then multicast via the RTP/UDP/IP stack to establish on-line interactions between server (instructor) and clients (students). A stream from the encoder is sent to the Single-Layer Feed Forward Controller to:
1. Dynamically characterise the nature of the video sequence (Temporal Information and Spatial Information). (Discussed in Section 6.3)

2. Use linear prediction for predicting the complexity of the incoming movie sequence. While the predicted value of the complexity is being used by the controller, the actual values are stored. (Discussed in Section 6.4)

3. Take input of the availability of the bandwidth.

4. Decide the different possible ways of configuration (frame rate and resolution) of the codec for the incoming video sequence. (Discussed in Section 6.6)

5. Select the best option for the optimum transmission of video. (Discussed in Section 6.7)

6. Send the optimised values of configuration parameters to the encoder to configure the codec for coming scene.

**Figure 6.1** schematically shows the flow diagram of the working of the controller.
6.3 DYNAMIC CHARACTERISATION OF THE ONGOING VIDEO SEQUENCE

The characterisation of the video sequences is done by calculating the Spatial Information (SI) and the Temporal Information (TI) values of every frame as recommended by the ITU-T P910 (09/99). This has been discussed in the Section 3.11.

For the calculation, ITU recommends that the calculations be performed on a sub-image of the video frame to avoid unwanted edge effects. At the frame rate of 25 fps each frame gets less than 0.04 s for the characterisation and optimisation. An ordinary PC (2 MB cache, 3 GHz clock speed) requires 0.47s for complete calculation for characterisation of a frame. Hence, to save on the computation time, a sub-image has to be used. Figures 6.2-6.4 show the effect of window size on the SI and TI results for the three sequences. The figures show that with the decrease in the size of the window, the difference between the values obtained by taking the whole frame and the values obtained by taking a section of the frame increases. Comparison between the results obtained for the whole image and the subsections show that by taking only 50% window size, i.e. by leaving 75 pixel from top and bottom, it is possible to obtain an acceptably good match of the values of SI and TI for the whole frame and part of the frame, i.e. the error in using such a subsection, would be less than 10% for SI and 20% for TI. Further saving on the calculation time may be achieved by efficient programming and use of a faster computer.

The selection of sub-image is done judiciously to allow fast computation and at the same time be representative of the complete frame. For example, in the case of panorama shot, most of the frame is occupied by non-moving inanimate objects such as blackboard, table, etc. There is only a small area where the instructor can move. Hence, only that area is selected for the calculation. This allows the reduction in calculation time.
FIGURE 6.2. Effect of window size on the calculation of SI and TI for the Head-and-Shoulder sequence.
FIGURE 6.3. Effect of window size on the calculation of SI and TI for the Middle sequence.
FIGURE 6.4. Effect of window size on the calculation of SI and TI for a sequence.
6.4 PREDICTION OF THE QUALITY OF INCOMING VIDEO FRAMES

The controller has to predict the quality of incoming frames and then optimise the codec configuration. The raw video is captured at the rate of 25 frames per second, hence, the SI and TI values of twenty five frames are calculated. Using the concept of linear regression analysis, the standard least squares fitting technique, the best-fit linear curve for the twenty five points is found. This equation is then used to predict the expected values of the thirty-seventh frame. The thirty seventh frame was selected to predict the average value of the next twenty five frames.

Figure 6.5 shows the SI and TI values of the three sequences – Head-and-Shoulder, Middle and Panorama – and the corresponding predicted values. The figure shows a reasonably good match between the actual and predicted SI and TI values if 25 frames are used for the calculation. Only during the change of scenes do the predicted values overshoot unreasonably. This overshoot is more drastic in the case of TI than for SI. In order to take care of this overshoot a condition was introduced according to which, if the predicted TI value was far beyond the previously calculated value, then the predicted value would be equal to the previous value (Figure 6.6). This assumption was considered as reasonable under the conditions of the video sequences. As a result of this additional clause, the actual SI and TI values and the predicted SI and TI values matched reasonably well.

To allow for the codec to implement the recommendations of the controller, the optimisation calculations are done every second. Hence, in the controller, twenty-five frames are used to predict the average SI and TI values of the next twenty five frames. While the predicted values are being used by the controller, the actual values are stored and used to calculate the next predicted values. Under such conditions, while there is a reasonable good match between the actual and predicted SI values, the match for TI is not so close (Figure 6.7).
FIGURE 6.5. Complexity (SI and TI) of a video sequence with overshoots.
FIGURE 6.6. Complexity (SI and TI) of a video sequence without overshoots.
FIGURE 6.7. Complexity (SI and TI) of a video sequence as used by the controller.
6.5 DEVELOPMENT OF THE CONTROLLER EQUATIONS

6.5.a Relationship between Bitrate – Frame Rate – Resolution

The MPEG-4 codec allows encoding at different frame rates and frame sizes. In order to vary the spatial resolution, the video sequence was first sub-sampled to the lower size, while maintaining the aspect ratio, and then up-sampled to the original size. For example, in order to get 50% resolution, the original image having the size 352 x 288 pixels, was first sub-sampled to 176 x 144 pixels and then up-sampled back to 352 x 288 pixels.

Thus, by varying the frame rate from 5 to 25 fps in the steps of 5 fps, and by varying from 10 to 100% in the steps of 10%, a total of 50 different quality encoded sequences of each standard sequence were obtained. The average bitrate of each of these encoded sequences was then taken. Figures 6.8-6.10 show the plot of Bitrate – Frame Rate – Resolution for the three sequences. As is to be expected, the Bitrate increases with an increase in the Frame Rate and Resolution. The figures also show that:

\[ \text{Bitrate} \propto \text{Frame Rate} \]
\[ \text{Bitrate} \propto (\text{Resolution})^n \]

Since, resolution is a two-dimensional (length and breadth) parameter, it is reasonable to presume that Bitrate would be proportional to the square of resolution.

In the second step, best-fit curve of Bitrate – Frame Rate – Resolution was plotted for the sequences. These best-fit curves were calculated using SigmaPlot Version 8. Hence, a general equation involving Frame Rate, Resolution and Resolution$^2$ was considered for modeling. Hence, the best-fit curves were modelled according to the form:

\[
\text{Bit Rate} = A + B \times (\text{Frame Rate}) + C \times (\text{Resolution}) + D \times (\text{Resolution}^2) + E \times (\text{Frame Rate} \times \text{Resolution}) + F \times (\text{Frame Rate} \times \text{Resolution}^2)
\]  

(6.1)

Figures 6.11-6.13 show the experimental values and the best-fit curves according to the Equation 6.1 for the three sequences. The figures show a close match between the experimental values and the best-fit curves. Table 6.1 shows the coefficient of the curves for the three sequences. Similar curves were plotted for seven more sequences and their coefficients were tabulated.
FIGURE 6.8. Effect of Frame Rate and Resolution on the bandwidth requirement for the Head-and-Shoulder sequence.

FIGURE 6.9. Effect of Frame Rate and Resolution on the bandwidth requirement for the Middle sequence.

FIGURE 6.10. Effect of Frame Rate and Resolution on the bandwidth requirement for the Panorama sequence.
\textbf{FIGURE 6.11.} Best-fit curve for the Head- and Shoulder sequence.

Bitrate = 5.5170 - 0.5480 \times FR - 0.2960 \times Res + 0.0056 \times Res^2
+ 0.0519 \times FR \times Res - 0.0004 \times FR \times Res^2

\textbf{FIGURE 6.12.} Best-fit curve for the Middle sequence.

Bitrate = 7.3543 - 0.4202 \times FR - 0.3728 \times Res + 0.0055 \times Res^2
+ 0.0362 \times FR \times Res - 0.0003 \times FR \times Res^2

\textbf{FIGURE 6.13.} Best-fit curve for the Panorama sequence.

Bitrate = 9.7116 - 0.8424 \times FR - 0.4852 \times Res + 0.0074 \times Res^2
+ 0.0619 \times FR \times Res - 0.0005 \times FR \times Res^2
TABLE 6.1. Comparison between the coefficients of the three sequences found from best-fit curves and generalised equations.

<table>
<thead>
<tr>
<th></th>
<th>Best-fit Curves</th>
<th>Generalised Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Head-and-Shoulder</td>
<td>Middle</td>
</tr>
<tr>
<td>SI</td>
<td>52.6</td>
<td>38.8</td>
</tr>
<tr>
<td>TI</td>
<td>4.4</td>
<td>2.9</td>
</tr>
<tr>
<td>B</td>
<td>-0.54798</td>
<td>-0.42017</td>
</tr>
<tr>
<td>C</td>
<td>-0.29604</td>
<td>-0.37281</td>
</tr>
<tr>
<td>D</td>
<td>0.005552</td>
<td>0.005523</td>
</tr>
<tr>
<td>E</td>
<td>0.051895</td>
<td>0.036236</td>
</tr>
<tr>
<td>F</td>
<td>-0.00044</td>
<td>-0.00025</td>
</tr>
</tbody>
</table>

Reference Figure 6.11 Figure 6.12 Figure 6.13 Figure 6.20 Figure 6.21 Figure 6.22

6.5b Relationship between Bitrate – Spatial Information – Temporal Information

Having got all the values of the coefficients for the video sequences, a common equation was developed linking Bitrate – Frame Rate – Resolution with the characteristics of the video sequences. This was done by expressing the coefficients A-F in terms of the Spatial Information and Temporal Information. It is reasonable to expect that the coefficients of terms including frame rate are functions of the temporal information and the coefficients of the terms including resolution are functions of the spatial information. By plotting the values of the coefficients against the SI and TI values and curve fitting, correlations were obtained. Figures 6.14-6.19 show the correlation between A to F and SI and TI.

\[
\text{Bit Rate} = A + B \times (\text{Frame Rate}) + C \times (\text{Resolution}) + D \times (\text{Resolution}^2) + E \times (\text{Frame Rate} \times \text{Resolution}) + F \times (\text{Frame Rate} \times \text{Resolution}^2)
\]  
\[ (6.1) \]

\[
A = 1.5829 \times TI \quad \text{(Figure 6.14)} \quad (6.2)
\]
\[
B = -0.1325 \times TI \quad \text{(Figure 6.15)} \quad (6.3)
\]
\[
C = -0.2531 - 0.0027 \times SI \quad \text{(Figure 6.16)} \quad (6.4)
\]
\[
D = 0.0024 + 0.00008 \times SI \quad \text{(Figure 6.17)} \quad (6.5)
\]
\[
E = 0.004623 \times TI + 0.000595 \times SI \quad \text{(Figure 6.18)} \quad (6.6)
\]
\[
F = 0.000298 + 0.0000118 \times TI - 0.000015 \times SI \quad \text{(Figure 6.19)} \quad (6.7)
\]

FIGURE 6.15. Best-fit curve for coefficient B.

FIGURE 6.16. Best-fit curve for coefficient C.

FIGURE 6.17. Best-fit curve for coefficient D.

FIGURE 6.18. Best-fit curve for coefficient E.

Substituting the values of A to F by their respective correlations, general correlations between Bitrate – Temporal Information – Spatial Information – Frame rate – Resolution were developed. Figures 6.20-6.22 show examples of this calculation for the three sequences. The figures show a good match between the experimental values, best-fit curve and the results obtained using the generalised equations between Bitrate – SI – TI – Frame Rate – Resolution. (Equations 6.1-6.7). Table 6.1 shows the comparison between the coefficients of the three sequences found from best-fit curves and generalised equations. The table shows that the values match quite closely.
75

Chapter 6

Design of Single-Layer Feed-Forward Controller

FIGURE 6.22. Experimental values, best-fit curve and values calculated curve using the generalised equation for Panorama sequence (SI = 55.1; T1 = 6.3).

Best-fit curve:

- Bitrate = 9.7116 - 0.8424 x FR - 0.4852 x Res + 0.0074 x Res² + 0.0619 x FR x Res - 0.0005 x FR x Res²

Calculated:

- Bitrate = 9.9723 - 0.8348 x FR - 0.4019 x Res + 0.0068 x Res² + 0.0619 x FR x Res - 0.0005 x FR x Res²

6.5.c Relationship between Perceptual Quality – Frame rate – Resolution

Extensive work has been done by Cranley et al. [2003, 2004] to develop correlation between Perceptual Quality – Frame rate – Resolution. They have developed a general relationship for the Optimum Adaptation Trajectory which defines the correlation between the Frame rate and Resolution, which in turn gives the optimum perceptual quality for the given bitrate. The equation developed by them has been used for the development of the controller. According to them:

\[
\text{Resolution} = 21 \ln (\text{Frame Rate}) + 32
\]

(6.8)

Table 6.2 shows the calculated values for the Equation 6.8.
**TABLE 6.2. Relationship between the frame rate and resolution for the Optimum Adaptation Trajectory [Cranley et al. 2003, Cranley 2004]**

<table>
<thead>
<tr>
<th>Frame rate (fps)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (%)</td>
<td>32</td>
<td>47</td>
<td>55</td>
<td>61</td>
<td>66</td>
<td>70</td>
<td>73</td>
<td>76</td>
<td>78</td>
<td>80</td>
<td>82</td>
<td>84</td>
<td>86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frame rate (fps)</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (%)</td>
<td>87</td>
<td>89</td>
<td>90</td>
<td>91</td>
<td>93</td>
<td>94</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>98</td>
<td>99</td>
<td>100</td>
</tr>
</tbody>
</table>

Even though the equation was developed for the video sequences provided by VQEG (Appendix 2), it has been assumed that it would satisfy the requirements of the sequences used for this research also. It is accepted that since this equation was developed using the subjective tests for a different set of sequences it is not correct to apply it here. Similar equations can be developed for the sequences used for this research.

### 6.6 CALCULATION OF THE AVAILABLE OPTIONS

Figure 6.23 shows an example of the decision-making process of the controller. In the first step the controller characterises the ongoing video sequence in terms of SI / TI and then predicts the complexity of the incoming sequence. Based on the predicted SI / TI values it is possible to calculate the values of A-F according to the Equations 6.2-6.7. From these values of A-F it is possible to calculate the bandwidth requirement for different frame rates and resolutions using the Equation 6.1. Figure 6.23 shows the Bitrate – Frame rate – Resolution curve for the predicted characteristics of the incoming sequence.

At the next step, the controller then reads the value of available bandwidth. For example, if the feedback merger informs the availability of 20 kbps, then it is possible to transmit video in 25 different ways, for example (frame rate, resolution) = (5, 77%), (10, 67%), (15, 54%), (20, 42%) and (25, 35%). In Figure 6.23 the intersection of the – Frame rate – Resolution curve with the available bandwidth plane gives the 25 different options available to the controller for transmitting the incoming sequence within the available bandwidth.
6.7 SELECTION OF THE BEST OPTION FROM THE AVAILABLE OPTIONS

Out of these 25 ways the video sequence can be transmitted within the available bandwidth, the best perceptual quality is obtained only by one combination of Frame rate and Resolution. This option is calculated using the equation for the Optimum Adaptation Trajectory (Equation 6.8). Figure 6.23 shows that the three equations (a) Bitrate – Frame rate – Resolution, (b) Available Bandwidth, and (c) Optimum Adaptation Trajectory intersect at one combination of Frame rate and Resolution. This combination represents the combination for the incoming video sequence which when transmitted within the available bandwidth will give perceptually the best quality of video at the receiving end. For the example shown in the Figure 6.23, the best combination is the one having (Frame rate, Resolution) = (7, 73%).

The decision-making process for the Single-Layer Feed Forward Controller can be summarised as:

1. Read the available bandwidth at the receiver end.
2. Store the SI and TI values of the ongoing 25 frames of the video sequence.
3. Use linear prediction for calculating the expected SI and TI values of the incoming movie sequence using the previous 25 values.
4. Use the predicted SI and TI values to calculate the coefficient A to F that characterises the calculated Frame rate – Resolution – Bitrate of the video sequence.
5. Calculate the 25 available options of (Frame rate, Resolution) by varying the value of frame rate from 1 to 25 and bitrate equal to the available bitrate.
6. Calculate the optimum combination from these 25 options using the equation for the Optimum Adaptation Trajectory.
FIGURE 6.23. Working of the controller in the case of the Head-and-Shoulder sequence. For 20 kbps it is possible to have 25 options, which is obtained by the intersection of the two curves Bit Rate – Frame Rate – Resolution and Available Bandwidth (20 kbps). Out of these 25 options there is one condition Frame Rate = 7 fps, Resolution = 73% which will give the best perceptual quality which is obtained by the intersection of the three curves Bit Rate – Frame Rate – Resolution, Available Bandwidth (20 kbps) and the optimum Adaptation Trajectory.
6.8 SUMMARY AND CONCLUSIONS

This chapter describes the work done for finding scheme for transmitting the video sequence within the available bandwidth in a perceptually optimised way. It describes the results of the experiments carried out to find the bandwidth requirement for transmitting different sequences at different frame rate and resolution. It also describes that for a video sequence, the values of Bitrate – Frame rate – Resolution can be satisfied using best-fit curve of the type

\[
\text{Bitrate} = A + B \times (\text{Frame Rate}) + C \times (\text{Resolution}) + D \times (\text{Resolution}^2) + E \times (\text{Frame Rate} \times \text{Resolution}) + F \times (\text{Frame Rate} \times \text{Resolution}^2)
\]

The values of the coefficients A-F can be related to SI and TI values of the sequence using empirical correlations. Hence, a generalised correlation between Bitrate – Frame rate – Resolution – SI – TI has been empirically found for the test video sequences.

Based on these equations, a Single-Layer Feed-Forward Controller has been proposed for transmitting video sequences in a perceptually optimised way within the available bandwidth.

This concept needs to be evaluated in order to judge the suitability of the proposed scheme. The evaluation of the Single-Layer Feed-Forward Controller is discussed in the next chapter, Chapter 7.
CHAPTER 7

EVALUATION OF THE SINGLE-LAYER FEED-FORWARD CONTROLLER
7.1 INTRODUCTION

It has been proposed in this research work that it may be possible to efficiently utilise the available bandwidth for the transmission of e-learning videos by dynamically adjusting the frame rate and resolution of the video sequences. In the previous chapter, Chapter 6, the concept of Single-Layer Feed-Forward Controller to dynamically control the frame rate and resolution of the transmitted video has been discussed. This controller needs to be evaluated for its performance in order to prove the hypothesis. This chapter deals with the evaluation of the Feed-Forward Controller.

In order to prove the viability of the controller, the performance of the controller needs to be evaluated. This can be done either in a real network or by simulation. Unfortunately, the evaluation could not be done in the real network due to the following reasons:

- The controller takes 0.47 s for the characterisation of each frame, whereas only 0.04 s is actually available to it (see Section 6.3) for the task.
- The MPEG4 codec used for the experiments (FFMPEG) is not properly documented, hence it is difficult to find the proper sections for the implementation of the controller and to use it for real-time transmission.

Hence, the performance of the controller could only be evaluated in a simulated environment.

To study the behaviour of the controller, a three-dimensional matrix of the variables was prepared (Figure 7.1). The three variables of the matrix were:

- Presence of the controller – yes or no.
- Availability of bandwidth – low or high.
- Type of available bandwidth – constant or variable.

Table 7.1 shows the eight conditions that were studied for the evaluation of the controller.
FIGURE 7.1. Matrix for the evaluation of the controller.

TABLE 7.1. Matrix for the evaluation of the controller.

<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>Presence of Controller</th>
<th>Availability of Bandwidth</th>
<th>Type of Available Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>Low</td>
<td>Constant</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Low</td>
<td>Variable</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>High</td>
<td>Constant</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>High</td>
<td>Variable</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Low</td>
<td>Constant</td>
</tr>
<tr>
<td>6</td>
<td>Yes</td>
<td>Low</td>
<td>Variable</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>High</td>
<td>Constant</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>High</td>
<td>Variable</td>
</tr>
</tbody>
</table>
Table 7.2. Relationship between the frame rate and resolution for the Optimum Adaptation Trajectory [Cranley 2003].

<table>
<thead>
<tr>
<th>Frame rate (fps)</th>
<th>Resolution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
</tr>
<tr>
<td>3</td>
<td>61</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>73</td>
</tr>
<tr>
<td>7</td>
<td>76</td>
</tr>
<tr>
<td>8</td>
<td>78</td>
</tr>
<tr>
<td>9</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>82</td>
</tr>
<tr>
<td>11</td>
<td>84</td>
</tr>
<tr>
<td>12</td>
<td>86</td>
</tr>
<tr>
<td>13</td>
<td>87</td>
</tr>
<tr>
<td>14</td>
<td>89</td>
</tr>
<tr>
<td>15</td>
<td>90</td>
</tr>
<tr>
<td>16</td>
<td>91</td>
</tr>
<tr>
<td>17</td>
<td>93</td>
</tr>
<tr>
<td>18</td>
<td>94</td>
</tr>
<tr>
<td>19</td>
<td>95</td>
</tr>
<tr>
<td>20</td>
<td>96</td>
</tr>
<tr>
<td>21</td>
<td>97</td>
</tr>
<tr>
<td>22</td>
<td>98</td>
</tr>
<tr>
<td>23</td>
<td>99</td>
</tr>
<tr>
<td>24</td>
<td>100</td>
</tr>
<tr>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

7.2. EXPERIMENTAL PROCEDURE

7.2.a Video Sequence

A test video sequence of 24 seconds was prepared by joining three 8-second sequences: (a) Head-and-Shoulder, (b) Middle and (c) Panorama. This test sequence thus had all the three important scenarios.

7.2.b Frame Rates and Resolution of the Test Sequences

Since the frame rate can be varied between 1 fps and 25 fps, and the resolution can be varied between 1% and 100%, there are 2500 ways (25 frame rates x 100 resolutions) of transmitting a video sequence. It is not possible to characterise all of these sequences and use them for the study. Hence, only twenty five frame rate – resolution pairs, as recommended by Cranley et al. [2003, 2004], were characterised and studied (Table 7.2).

7.2.c Availability of Bandwidth

Figure 7.2 shows the bandwidth requirement for the test sequence, which has been calculated using Equations 6.1-6.7. The figure shows that the bandwidth requirement varies between 3-113 kbps, and the average is 46 kbps.
Constant Bandwidth

For the evaluation of the performance of the controller, two constant bandwidths, 23 kbps and 31 kbps — half and two-thirds of the average required bandwidth — were selected. These values were selected arbitrarily. Testing at higher bandwidths would allow a significant section being transmitted at highest resolution and frame rate (100%, 25fps) in both — controlled and uncontrolled — conditions, thereby making the judgement of the performance difficult. Testing at very low bandwidths would result in very poor video quality, which would not be suitable for testing.

Variable Bandwidth

In order to get representative bitrate variations, a large file was downloaded from the University of Leeds, U.K. The available bitrate at the receiver end was then recorded using a freely-available software — NetMeter [NetMeter Homepage 2006]. Figure 7.3 shows the availability of bandwidth as a function of time. It is accepted that the transmission of real-time video is on UDP and the data transfer takes place over TCP, hence, the trend shown in Figure 7.3 cannot be used directly. At this stage it has been considered acceptable because of two reasons:

FIGURE 7.2. Bandwidth requirement for the test sequence.
The purpose is just to get the range of bandwidth availability. This value compares well with the report from Microsoft which has described the average bandwidth usage by NetMeeting. According to the report, NetMeeting uses (a) less than 50 kbps for a typical data conferencing scenario between two NetMeeting users (application sharing, whiteboard, chat, and file transfer); (b) less than 10 kbps for one-way audio; and (c) less than 40 kbps for a typical audio and video conferencing scenario (average movement using a medium-sized video window with medium quality) between two NetMeeting users [Microsoft 2006].

In TCP the traffic adjusts itself to the available bandwidth, hence, there is a relationship between the available bandwidth and the rate of transfer. By monitoring the rate of transfer it is possible to monitor the available bandwidth. On the other hand, the transfer rate in UDP is independent of the available bandwidth; hence it is not possible to correlate the available bandwidth by monitoring a real-time video transfer rate.

The evaluation of the variable bandwidth was carried out at 17% and 12% of the available bandwidth because at these two values the average available bandwidth was 30 kbps and 22 kbps, which were comparable to the 30 kbps and 23 kbps set for the constant bandwidth (Figure 7.4).
In this research work, the effectiveness of the Feed Forward Controller has been judged by the efficiency of the utilisation of the available bandwidth. The effective utilisation of the bandwidth has only been quantified by calculating under and over subscription of the bandwidth (Figure 7.5). It has been assumed that the over-utilisation of the bandwidth leads to congestion and packet drops, and the under utilisation would result in the transmission of poor quality video. Hence, the optimum subscription is when there is minimum over- and under-utilisation of the available bandwidth.

7.2.d Simulation of the Output Sequence

Since it was not possible to run the controller real-time in the available codec, the output from the controller had to be simulated. In order to do this, the controller was run separately for the pre-determined bandwidth and the recommendations made by the controller were recorded. The video sequences were then manually prepared by joining sections prepared at the frame rate and resolution as recommended by the controller.
7.2.e Objective Test

The quality of the test video sequence was tested objectively by four different methods:

- PSNR
- S1 based on the recommended of the ITU-T P910 (09/99)
- T1 based on the recommended of the ITU-T P910 (09/99)
- Feng Xiao’s DCT-based VQM.

For the test, the MPEG-4 sequences were converted to uncompressed AVI at 25 fps, using a freely available software – VirtualDub. This was done because in objective tests, individual frames are compared against corresponding reference frames pixel by pixel, and if there is a difference between the frame rate and the frame size between the reference and the test sequences, then the tests will not run properly. For the calculation of PSNR and Feng Xiao’s DCT-based VQM, the 25fps and 100% resolution sequence was used as a reference sequence.

7.2.f Subjective Test

The perceptual quality of the video sequences was tested subjectively according to the International Telecommunications Union (ITU), Recommendation P.910. This standard was considered to be appropriate in the absence of any other standard made specifically for the conferencing tools.

The test was carried out according to the recommendations for the Degradation Category Rating (DCR), also called the Double Stimulus Impairment Scale method. In this test, the test sequences are presented in pairs:

- The first stimulus presented in each pair is always the reference.
- The second stimulus is the same source presented through one of the systems under test.

Fifty individuals (nineteen males, thirty one females) in the age group 18-45 years participated in this experiment. All had normal or corrected-to-normal vision. All participants were fully comfortable with the location and adapted to the prevailing illumination conditions before beginning their task. All participants took part in all three video sequences and the order of the experiment was randomised. The subjects were asked to consider the MPEG-4 video at 25fps and 100% resolution as the reference sequence, and all other video sequences were judged in comparison to this sequence. Audio was muted in the experiment because it has been reported that the quality of audio affects the subjective test of the video.
As recommended by the standard, the following five-level scale for rating overall quality was used:

5 Imperceptible
4 Perceptible but not annoying
3 Slightly annoying
2 Annoying
1 Very annoying

7.2. g Experiments

Eight experiments were carried out according to Table 7.3. In four of the eight experiments the controller was not used, and in the other four it was used.

For experiments 1-4, where the controller was not used, the frame rate - resolution pair was maintained constant throughout the test sequence. Figure 7.6 shows the requirement of bandwidth for the transmission of the test sequence at different frame rates and the corresponding resolution. Based on this figure the frame rates and resolutions were decided for experiments 1-4. In all the experiments the aim was to maintain the average utilisation as close to the average available bandwidth as possible

Experiment 1. The frame rate and resolution was maintained at 8fps and 76% because the average bandwidth requirement for this quality of transmission was 21kbps, which was close to the available 23 kbps.

Experiment 2. The available bandwidth varied (12% of the total available), but the average bandwidth was 21 kbps. Hence, the frame rate and resolution was maintained at 8fps and 76% because the average bandwidth requirement for this quality of transmission is 21 kbps, which was close to the available 23 kbps

Experiment 3. The frame rate and resolution was maintained at 13fps and 86% because the average bandwidth requirement for this quality of transmission was 31 kbps, which was close to the available 31 kbps.

Experiment 4. The available bandwidth varied (17% of the total available), but the average bandwidth was 30 kbps. Hence, the frame rate and resolution was maintained at 13 fps and 86% because the average bandwidth requirement for this quality of transmission was 30 kbps, which was close to the available 31 kbps
TABLE 7.3. Summary of the experiments carried out to evaluate the performance of the controller.

<table>
<thead>
<tr>
<th>Expt. Number</th>
<th>Controller Used</th>
<th>Bandwidth Availability</th>
<th>Type of Bandwidth</th>
<th>Frame Rate, fps</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>23 kbps</td>
<td>Constant</td>
<td>8</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>12% of available</td>
<td>Variable</td>
<td>8</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average = 23 kbps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>31 kbps</td>
<td>Constant</td>
<td>13</td>
<td>86</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>17% of available</td>
<td>Variable</td>
<td>13</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average = 32 kbps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>23 kbps</td>
<td>Constant</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Yes</td>
<td>12% of available</td>
<td>Variable</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average = 23 kbps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>31 kbps</td>
<td>Constant</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>17% of available</td>
<td>Variable</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average = 32 kbps</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For experiments 5-8, where the controller was used, the frame rate - resolution pairs were varied so as to maintain the transmission just below the available bandwidth.

**Experiment 5.** The available bandwidth was 23 kbps and the controller recommended the adjustment of the frame rate – resolution pair as shown in Figure 7.7.

**Experiment 6.** The available bandwidth varied (12% of the total available) as shown in Figure 7.4. The recommendation made by the controller is shown in Figure 7.8.

**Experiment 7.** The available bandwidth was 31 kbps and the controller recommended the adjustment of the frame rate – resolution pair as shown in Figure 7.9.

**Experiment 8.** The available bandwidth varied (17% of the total available) as shown in Figure 7.4. The recommendation made by the controller is shown in Figure 7.10.
FIGURE 7.6. Average bandwidth required for transmission of the test sequence at different frame rates.
FIGURE 7.7. Recommendations made by the controller for Experiment 5. The various portions of the graph show (a) The actual and predicted TI values, (b) The actual and predicted SI values, (c) Resolution as calculated by the controller, (d) Frame rate as calculated by the controller, (e) available bandwidth and used bandwidth.
FIGURE 7.8. Recommendations made by the controller for Experiment 6. The various portions of the graph show (a) The actual and predicted TI values, (b) The actual and predicted SI values, (c) Resolution as calculated by the controller, (d) Frame rate as calculated by the controller, (e) Available bandwidth, used bandwidth and required bandwidth for 25 fps and 100% resolution.
FIGURE 7.9. Recommendations made by the controller for Experiment 7. The various portions of the graph show (a) The actual and predicted TI values, (b) The actual and predicted SI values, (c) Resolution as calculated by the controller, (d) Frame rate as calculated by the controller, (e) available bandwidth and used bandwidth.
FIGURE 7.10. Recommendations made by the controller for Experiment 8. The various portions of the graph show (a) The actual and predicted TI values, (b) The actual and predicted SI values, (c) Resolution as calculated by the controller, (d) Frame rate as calculated by the controller, (e) available bandwidth, used bandwidth and required bandwidth for 25 fps and 100% resolution.
7.3 ASSESSMENT OF THE OBJECTIVE TESTS

The objective tests that were used for evaluation of the perceptual quality of the video sequences had to be tested for their suitability to the task.

Twenty-five sequences were prepared by adjusting the frame rate and resolution of the test sequence according to Table 7.2. These twenty-five sequences were then characterised frame-by-frame using (a) PSNR values; (c) Spatial Information (SI), (d) Temporal Information (TI) and (b) Feng Xiao’s VQM value. The effect of frame rate on the average test values of all the frames of the test sequence (600 in number) was also studied. For the calculations of PSNR and Feng Xiao’s VQM values, the sequence having a frame rate of 25 fps and 100% resolution was used as the reference sequence.

7.3.1 PSNR

Figure 7.11 shows the PSNR values of the Y component of the sequences having frame rates of 10fps and 25fps for the 600 frames of the sequences. The figure shows that the PSNR values for the sequence prepared at 24fps are less than the PSNR values for the sequence prepared at 10fps. Figure 7.12 shows the effect of frame rate on the average PSNR values for the 25 sequences. The figure shows that there is no particular trend of the results. Thus, Figures 7.11 and 7.12 show that the PSNR values cannot be used for characterising the quality of the video sequences.

**FIGURE 7.11.** PSNR values of different frames for Y component of sequences prepared at 10 fps and 24 fps.

**FIGURE 7.12.** Average PSNR values of sequences prepared at different frame rates.
7.3. **Spatial Information (SI)**

Figures 7.13 and 7.14 show that the Spatial Information (SI) increases with the increase in the quality (frame rate and resolution) of the sequences. Hence, SI can be used for the characterisation of the test sequences.

7.3. **Temporal Information (TI)**

Figures 7.15 - 7.17 show that the Temporal Information (TI) increases with the increase in the quality (frame rate and resolution) of the sequences. Hence, TI can be used for characterisation of the test sequences.

Figure 7.16 shows that the sequence made at 24 fps has small changes between the two consecutive frames. On the other hand, for the sequence made at 10 fps, the TI values alternate between a large value and 0, implying that there are same 2-3 consecutive frames followed by a big change.
FIGURE 7.15. Temporal Information of different frames of sequences prepared at 10 fps and 24 fps.

FIGURE 7.16. Temporal Information of different frames of sequences prepared at 10 fps and 24 fps for the first 50 frames.

FIGURE 7.17. Average Temporal Information of sequences prepared at different frame rates.
7.3.4 Feng Xiao’s VQM

Figures 7.18 - 7.19 show that the Feng Xiao’s VQM steadily decreases with an increase in the quality (frame rate and resolution) of the sequences. Hence, Feng Xiao’s VQM can be used for characterisation of the test sequences.

From the study carried out it may be concluded that while PSNR cannot be used, SI, TI and Feng Xiao’s VQM can be used for characterising video sequences objectively. Since there are complicated relationships between the objective test values and the quality of the video sequences, the results should only be taken qualitatively and not quantitatively.

7.4 RESULTS AND DISCUSSION

7.4.1 Quality of Video

The perceptual quality of the video sequences prepared for the eight experiments was characterised using objective tests and subjective tests. Tables 7.4 and 7.5 show the objective and subjective test for the eight experiments.
### TABLE 7.4. Video quality of the sequences produced in the eight experiments.

<table>
<thead>
<tr>
<th>Expt.</th>
<th>Bandwidth Availability</th>
<th>Frame rate /Resolution</th>
<th>Average Bandwidth Used</th>
<th>Average SI Value</th>
<th>Average TI Value</th>
<th>Average VQM Value</th>
<th>Subjective Test Value (MOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23 kbps</td>
<td>8fps, 76%</td>
<td>22 kbps</td>
<td>43.9</td>
<td>2.6</td>
<td>2.8</td>
<td>3.1</td>
</tr>
<tr>
<td>2</td>
<td>12% of available</td>
<td>8fps, 76%</td>
<td>22 kbps</td>
<td>43.9</td>
<td>2.6</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>3</td>
<td>31 kbps</td>
<td>13fps, 86%</td>
<td>31 kbps</td>
<td>45.3</td>
<td>3.2</td>
<td>2.2</td>
<td>3.2</td>
</tr>
<tr>
<td>4</td>
<td>17% of available</td>
<td>13fps, 86%</td>
<td>31 kbps</td>
<td>45.3</td>
<td>3.2</td>
<td>2.2</td>
<td>3.5</td>
</tr>
<tr>
<td>5</td>
<td>23 kbps</td>
<td>Fig. 6.6</td>
<td>21 kbps</td>
<td>43.6</td>
<td>2.6</td>
<td>2.8</td>
<td>2.9</td>
</tr>
<tr>
<td>6</td>
<td>12% of available</td>
<td>Fig. 6.7</td>
<td>20 kbps</td>
<td>43.5</td>
<td>2.7</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>7</td>
<td>31 kbps</td>
<td>Fig. 6.8</td>
<td>30 kbps</td>
<td>45.2</td>
<td>3.2</td>
<td>1.9</td>
<td>3.4</td>
</tr>
<tr>
<td>8</td>
<td>17% of available</td>
<td>Fig. 6.9</td>
<td>29 kbps</td>
<td>45.0</td>
<td>3.2</td>
<td>2.0</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 7.5. Subjective test of the video sequences produced in the eight experiments.

<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>Total votes</th>
<th>Imperceptible</th>
<th>Perceptible but not annoying</th>
<th>Slightly annoying</th>
<th>Annoying</th>
<th>Very annoying</th>
<th>MOS</th>
<th>Std</th>
<th>%IOP</th>
<th>%AOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expt. 1</td>
<td>50</td>
<td>0</td>
<td>18</td>
<td>27</td>
<td>5</td>
<td>0</td>
<td>3.1</td>
<td>0.63</td>
<td>36</td>
<td>10</td>
</tr>
<tr>
<td>Expt. 2</td>
<td>50</td>
<td>0</td>
<td>17</td>
<td>27</td>
<td>6</td>
<td>0</td>
<td>3.0</td>
<td>0.65</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>Expt. 3</td>
<td>50</td>
<td>2</td>
<td>25</td>
<td>17</td>
<td>6</td>
<td>0</td>
<td>3.2</td>
<td>0.76</td>
<td>54</td>
<td>12</td>
</tr>
<tr>
<td>Expt. 4</td>
<td>50</td>
<td>1</td>
<td>31</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>3.5</td>
<td>0.60</td>
<td>64</td>
<td>4</td>
</tr>
<tr>
<td>Expt. 5</td>
<td>50</td>
<td>0</td>
<td>14</td>
<td>29</td>
<td>7</td>
<td>0</td>
<td>2.9</td>
<td>0.64</td>
<td>28</td>
<td>14</td>
</tr>
<tr>
<td>Expt. 6</td>
<td>50</td>
<td>0</td>
<td>21</td>
<td>22</td>
<td>7</td>
<td>0</td>
<td>3.0</td>
<td>0.70</td>
<td>42</td>
<td>14</td>
</tr>
<tr>
<td>Expt. 7</td>
<td>50</td>
<td>3</td>
<td>22</td>
<td>22</td>
<td>3</td>
<td>0</td>
<td>3.4</td>
<td>0.71</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>Expt. 8</td>
<td>50</td>
<td>5</td>
<td>23</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>3.4</td>
<td>0.81</td>
<td>56</td>
<td>10</td>
</tr>
</tbody>
</table>

**MOS:** Mean Opinion Score  
**Std:** Standard Deviation  
**%IOP:** Percentage of Imperceptible or Perceptible but not annoying  
**%AOV:** Percentage of Annoying or Very Annoying
There are four pairs of sequences that were prepared without and with the use of the Feed Forward Controller. These were:

- Experiments 1 and 5
- Experiments 2 and 6
- Experiments 3 and 7
- Experiments 4 and 8

A comparative study of these four pairs of experiments show that there is no difference between the quality of videos produced with or without the controller when the bandwidth profile is the same. This observation is to be expected because in the four pairs of experiments, the bandwidth used is nearly the same in both (with and without the controller) the cases.

### 7.4.b Utilisation of Bandwidth

The main purpose of the controller is to optimally use the available bandwidth for the transmission of the e-learning video. The optimal usage implies that the bandwidth required for the transmission of the video at any particular instance should closely match the bandwidth available at that instance. If the bandwidth requirement is less than the bandwidth available, it will lead to under-utilisation of the bandwidth (Figure 7.5), resulting in the transmission at inferior quality. On the other hand, if the bandwidth requirement is more than the bandwidth available, it will lead to the over-utilisation of the bandwidth (Figure 7.5). Over-utilisation is generally worse than the under-utilisation because over-utilisation may result in congestion and packet drops, resulting in severe degradation of the video quality at the receiver end.

In this research project, the transmission of video through the network and the quality of the video at the receiver-end was not studied. The work was carried out only at the sender end. Hence, the transmission losses and the quality of video at the receiver end can only be guessed. For this purpose it is assumed that the greater the over-utilisation of the bandwidth, the greater are the transmission losses due to congestion and packet drops.

For the comparison, four pairs of sequences were prepared without and with the use of the Feed Forward Controller. The four test pairs were:

- Experiments 1 and 5
- Experiments 2 and 6
- Experiments 3 and 7
- Experiments 4 and 8
**TABLE 7.6. Utilisation of the bandwidth by the video sequences produced in the eight experiments.**

<table>
<thead>
<tr>
<th>Expt.</th>
<th>Bandwidth Availability</th>
<th>Frame rate / Resolution</th>
<th>Controlled/ Uncontrolled</th>
<th>Average Bandwidth Used</th>
<th>Under Utilisation, %</th>
<th>Over Utilisation, %</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23 kbps</td>
<td>8fps, 76%</td>
<td>Uncontrolled</td>
<td>22 kbps</td>
<td>10</td>
<td>26</td>
<td>Fig. 7.20</td>
</tr>
<tr>
<td>2</td>
<td>12% of available</td>
<td>8fps, 76%</td>
<td>Uncontrolled</td>
<td>22 kbps</td>
<td>9</td>
<td>35</td>
<td>Fig. 7.21</td>
</tr>
<tr>
<td>3</td>
<td>31 kbps Average = 23 kbps</td>
<td>13fps, 86%</td>
<td>Uncontrolled</td>
<td>31 kbps</td>
<td>11</td>
<td>27</td>
<td>Fig. 7.22</td>
</tr>
<tr>
<td>4</td>
<td>17% of available Average = 32 kbps</td>
<td>13fps, 86%</td>
<td>Uncontrolled</td>
<td>31 kbps</td>
<td>11</td>
<td>30</td>
<td>Fig. 7.23</td>
</tr>
<tr>
<td>5</td>
<td>23 kbps Fig. 7.7</td>
<td>Controlled</td>
<td>21 kbps</td>
<td>16</td>
<td>18</td>
<td>24</td>
<td>Fig. 7.20</td>
</tr>
<tr>
<td>6</td>
<td>12% of available Average = 23 kbps</td>
<td>Fig. 7.8</td>
<td>Controlled</td>
<td>20 kbps</td>
<td>18</td>
<td>24</td>
<td>Fig. 7.21</td>
</tr>
<tr>
<td>7</td>
<td>31 kbps Fig. 7.9</td>
<td>Controlled</td>
<td>30 kbps</td>
<td>15</td>
<td>23</td>
<td>21</td>
<td>Fig. 7.22</td>
</tr>
<tr>
<td>8</td>
<td>17% of available Average = 32 kbps</td>
<td>Fig. 7.10</td>
<td>Controlled</td>
<td>29 kbps</td>
<td>17</td>
<td>21</td>
<td>Fig. 7.23</td>
</tr>
</tbody>
</table>

The bandwidth utilisation was calculated according to the **Equations 6.1 - 6.7. Figures 7.20-7.23 and Table 7.6** show the utilisation of the available bandwidth for the eight experiments. A comparative study of Figures 7.20-7.23 and Table 7.6 shows that in all the four cases, the bandwidth over-utilisation is less for the controlled transmission as compared to the uncontrolled transmission. This satisfies the major criterion for rate-shaping, which is to regulate the transmission according to the availability.

One shortcoming of the controller is that at times it tends to overshoot the bandwidth requirement. It does this under two conditions:

- When there is a change in the scene (Frames 200 and Frame 400). Whenever there is a sudden change in the scene the TI value increases drastically, the controller then temporarily becomes instable.
- When there are large movements in a scene. This is again due to large changes in the TI values, resulting in the wrong predictions.

Overall, the use of controller helps to effectively regulate the e-learning video parameters – frame rate and resolution – so as to transmit the video sequences within the available bandwidth without any loss in the perceptual quality.
FIGURE 7.20. Utilisation of constant low bandwidth under uncontrolled (Expt. 1) and controlled (Expt. 5) conditions.

FIGURE 7.21. Utilisation of variable low bandwidth under uncontrolled (Expt. 2) and controlled (Expt. 4) conditions.

FIGURE 7.22. Utilisation of constant high bandwidth under uncontrolled (Expt. 3) and controlled (Expt. 7) conditions.

FIGURE 7.23. Utilisation of variable high bandwidth under uncontrolled (Expt. 4) and controlled (Expt. 8) conditions.
7.5 SUMMARY AND CONCLUSIONS

The performance of a Feed Forward Controller for controlling the transmission of MPEG4 video for e-learning applications has been tested in a pseudo-real test environment. The tests were carried out to study the benefit of using the controller for the transmission of video under four conditions of bandwidth – high and low; constant and variable.

It has been found that the PSNR is not suitable for the evaluation of the quality of the video sequences; on the other hand, SI, TI and Feng Xiao’s VQM have been found suitable for the task.

It has been found that the quality of test video sequences is the same for both – controlled and uncontrolled – transmission when the average bandwidth utilisation is nearly the same.

It has also been found that the transmission of the video sequences under controlled condition results in less over-utilisation of the bandwidth as compared to the transmission of the video sequences under uncontrolled condition. Since the over-utilisation of bandwidth may result in congestion and packet drops, it may be expected that the use of controller would improve the quality of video transmission.

Hence, it may be concluded that the Feed Forward Controller dynamically adjusts the frame rate and resolution of the transmitted video sequences according to the available bandwidth, in order to lower the over-utilisation of the bandwidth. While lowering the over-utilisation, it maintains the perceptual quality of the transmitted video.
CHAPTER 8

CONCLUSIONS AND FUTURE WORK
CHAPTER 8

CONCLUSIONS AND FUTURE WORK

8.1 CONCLUSIONS

E-learning is becoming increasingly popular among students who are unable to attend regular classes. E-learning lectures can be classified into two different categories based upon the interaction between the instructor and students; these are synchronous and asynchronous. In synchronous e-learning the instruction involves real-time delivery of audio, video and other media. Unfortunately, in synchronous e-learning the quality of video is substandard and significant work needs to be done to improve it. This improvement can be brought about either by improving the hardware, such as increasing the computer processing power and increasing the bandwidth of network, or by modifying the software, such as by changing the codec and transmission protocols.

An understanding of the human visual system can provide a sound basis for improving the quality of video in a synchronous e-learning tool. In this research it has been proposed that it is possible to significantly improve the quality of video at the receiver end by first reducing the amount of data to be transmitted, and then by transmitting the important data in a controlled way so as to provide the best perceptual quality. Hence, the key issue that has been addressed in this research is how to adapt a video transmission in order to maximise the resulting user-perceived quality.

Current research in human visual system has shown that the human eye can only see one section of a scene at a time – either in detail or in a coherent (stable) way. Humans cannot see more than one object in detail and in a coherent way at the same time; and also it can focus attention only on one event at a time [Rensink 2000a, 2000b, 2002]. It has been proposed that these limitations of human eyes be exploited to define “foreground” and “background”. Once these two have been identified, they may be separated using suitable techniques. The foreground can then be transmitted in a perceptually optimised way within the available bandwidth and the background can be transmitted less frequently and in relatively poorer quality.
While most of the current adaptation techniques address the problems of transmission taking into account only the network conditions, in this work both – the network conditions and the human psycho-physical perceptual system – were considered for optimisation. In this research project it has been proposed that the available bandwidth can be most efficiently utilised by dynamically adapting the quality of video, based on the availability of the bandwidth.

In order to vary the quality of video according to the availability of bandwidth, the use of two different types of Feed-Forward Controllers has been proposed. The Multi-Layer Feed-Forward Controller helps to satisfy the varied bandwidth requirements of multiple receivers in a heterogeneous environment. It dynamically characterises the ongoing video sequence for complexity (number of Macroblocks per frame), takes the available bandwidth from different receivers and then adjusts the number of layers, distribution of frames among different layers and the frame rate before transmitting the layered video stream.

The Single-Layer Feed-Forward Controller inputs the video and dynamically characterises it in terms of the Spatial Information and Temporal Information. Based on the video characteristics and the availability of bandwidth, the controller then adapts the video transmission in terms of the frame rate and resolution so as to provide a perceptually optimised video.

The suitability of the proposed Feed-Forward Controllers for the dynamic adaptation of e-learning video has been evaluated in a simulated environment. It has been found that the use of the Multi-Layer Feed-Forward Controller allows an optimised transmission of layered video so as to improve the perceptual quality of video at the receiving end. The Single-Layer Feed-Forward Controller allows adaptation of the video quality in terms of frame rate and resolution so as to lessen over-subscription of the available bandwidth while maintaining the perceptual quality of video.
8.2 CONTRIBUTIONS

This research makes the following novel contributions:

8.2.a Definition of “foreground” and “background” in an e-learning video

This research work defines “foreground” and “background” in the context of an e-learning video sequence. It does this by first examining the psychological basis for the identification of objects in terms of Critical Interest and Marginal Interest, and then shows that it may be possible to correlate “foreground” and “background” as the objects of Central Interest and Marginal Interest, respectively. Thus this research, for the first time, provides a scientific method for defining “foreground” and “background” objects in an e-learning video.

8.2.b Relationship between the “inattentional blindness” and the perceptual quality of video

This study has also shown that while watching an e-learning sequence, humans suffer from “inattentional blindness”, because of which they fail to notice the quality of the background (objects of Marginal Interest) while they are focussing attention on the foreground (objects of Critical Interest). Further, it has been shown that in an e-learning video sequence, as long as the quality of the foreground (spatial resolution) is maintained at a high level, some degradation in the quality of the background does not significantly lower the overall perceptual quality of the sequence.

8.2.c Relationship between the “change blindness” and the perceptual quality of video

It has been shown that while watching an e-learning video, humans suffer from “change blindness”. As a result, they are unable to notice changes in the objects of Marginal Interest. Hence, when two similar e-learning videos, differing in the objects of Marginal Interest, and interspaced by a different video are shown to the viewers, the viewers find it difficult to notice the differences. Hence, it is possible to separate foreground and background, and then transmit the foreground frames at a relatively higher frame rate as compared to the background frames. This separate transmission of foreground and background frames will result in a saving on the bandwidth. The slight anomalies, which may arise due to infrequent transmission of the background, may go unnoticed by the viewers due to change blindness.
8.2.d  Proposal for a Multi-Layer Feed Forward Controller (FFC) for Optimised Transmission of Video

In this research work a Multi-Layer Feed-Forward Controller for the layered transmission of video has been proposed. It has been proposed that the controller first characterise the ongoing video sequence for complexity (number of Macroblocks per frame) and then predict the complexity of the incoming video sequence. Using the values of available bandwidth for different receivers and the predicted complexity it calculates the layering scheme, number of layers, distribution of frames among different layers and frame rate, so as to provide the best perceptual quality video to a group of receivers in a heterogeneous environment.

The performance of the controller was tested in a laboratory setup. It was observed that the use of the controller improved the quality of video at the receivers’ end.

8.2.e  Relationship between Spatial Information (SI) - Temporal Information (TI) - Frame Rate - Resolution - Bitrate

The relationship between the quality of a video sequence (Spatial Information and Temporal Information), frame size, resolution and required bandwidth has been investigated. The investigation has shown that it is possible to develop a set of general equations correlating the quality of video sequence (SI and TI) with the transmission parameters (frame size and resolution) and the required bandwidth. It has been shown that for a video sequence, the values of Bitrate – Frame rate – Resolution can be correlated using a best-fit curve of the type

\[ \text{Bitrate} = A + B \times (\text{Frame Rate}) + C \times (\text{Resolution}) + D \times (\text{Resolution}^2) + E \times (\text{Frame Rate} \times \text{Resolution}) + F \times (\text{Frame Rate} \times \text{Resolution}^2) \]

The values of the coefficients A-F can be related to the SI and TI values of the sequence using empirical correlations.

8.2.f  Proposal for a Single-Layer Feed Forward Controller (FFC) for Optimised Transmission of Video

In this research work a Single-Layer Feed-Forward Controller for the transmission of video sequences in a perceptually optimal way has been proposed. It has been proposed that the Feed-Forward Controller first of all characterise the nature of an incoming video sequence (Temporal Information and Spatial Information) and then uses the correlations between the SI – TI – Frame rate – Resolution – Bitrate to calculate the various Frame rate – Resolution options available for the given bandwidth. Finally, the controller decides upon the optimal way of transmitting video by controlling the frame rate and resolution.
8.2.g Evaluation of the Feed Forward Controller (FFC)

The suitability of the Feed-Forward Controller has been evaluated in a simulated environment. It was observed that it is possible to control the transmission parameters (frame rate and resolution) of e-learning sequences so as to stream the video sequences with reduced over-subscription of the available bandwidth.

8.3 FUTURE WORK

This research has only touched the surface of the optimal transmission of video based on the network and human perceptual system. It is recommended that in future the following be carried out:

8.3.a Implementation of the Feedback System

In this investigation the feedback mechanism has not been implemented. This should be implemented so that the performance of the controller can be tested in real time.

8.3.b Improvement in the Prediction of SI and TI

The actual and predicted values of SI and TI matched well but not very closely. A better method needs to be developed.

8.3.c Reduction in the Computation

The whole concept of calculating SI and TI on a pixel to pixel basis is highly computationally intensive. A better method needs to be developed in order to reduce the amount and time of computation.

8.3.d Implementation in an Actual Codec

Due to large-scale computation, the time taken for the controller exceeded the 1/25 second (time between the two frames). Hence, the controller could not be implemented in an actual codec. Therefore, the controller needs to be implemented in an actual codec in order to test its strengths.

8.3.e Testing of the Controller in Actual Transmission

Since the controller could not be implemented in an actual codec, its performance could not be judged in a real transmission. Hence, the quality of video with and without controller needs to be tested in a real network.
8.3.f Separation of Foreground and Background

The efficient separation of foreground and background could not be completed within the limited time-frame of this research. Hence, more work needs to be carried out in order to separate the foreground and background more effectively.

8.3.g Transmission of Foreground and Background in Different VOPs

The MPEG-4 video codec that was used did not support the layering of video. Hence, the foreground and background need to be transmitted in different layers using a codec that supports layering.

8.3.h Transmission of Video on Different VOPs through Feed Forward Controller

The overall advantage of the use of feed forward controller combined with the efficient separation of foreground from the background needs to be investigated.
REFERENCES


References


213
References


100. Pan, J., Lin, C.-W., Gu, G. and Sun, M.-T., 2002. A Robust Video Object Segmentation Scheme with Prestored Background Information.


124. Simonson, M., 1999. Readings in Distance Education, Monograph, Fischler School of Education and Human Services, Nova Southeastern University.


References


220


221


222


APPENDIX 1

EDGE DETECTION
APPENDIX 1

EDGE DETECTION

Edges can be defined as the regions with strong intensity contrasts, i.e., there is a jump in the intensity from one pixel to the next, in an image. Hence, edge detection involves identifying and locating these sharp discontinuities in an image.

The first step in the edge detection is the convolution of a kernel over the image frame. The kernel itself is normally a set of two kernels. One kernel identifies the horizontal boundary and the other identifies the vertical boundary. Once these results of the convolution have been obtained, the square root of the sum of the squares of the results of the convolution, is taken in order to normalise the horizontal and vertical edge detections.

A number of kernels have been proposed. Some of the commonly used edge detectors are shown in Figure A1.1.

As an example, to find edge using Sobel filter:

\[
Gv(i, j) = -1 \times x(i - 1, j - 1) - 2 \times x(i - 1, j) - 1 \times x(i - 1, j + 1) + \\
+ 0 \times x(i, j - 1) + 0 \times x(i, j) + 0 \times x(i, j + 1) + \\
+ 1 \times x(i + 1, j - 1) + 2 \times x(i + 1, j) + 1 \times x(i + 1, j + 1)
\]  
(A1.1)

\[
Gh(i, j) = -1 \times x(i - 1, j - 1) + 0 \times x(i - 1, j) + 1 \times x(i - 1, j + 1) + \\
- 2 \times x(i, j - 1) + 0 \times x(i, j) + 2 \times x(i, j + 1) + \\
- 1 \times x(i + 1, j - 1) + 0 \times x(i + 1, j) + 1 \times x(i + 1, j + 1)
\]  
(A1.2)

\[
y(i, j) = \sqrt{[Gv(i, j)]^2 + [Gh(i, j)]^2}
\]  
(A1.3)
FIGURE A1.1 Kernels for various edge detectors.

Where:

$x(i, j)$ denote the pixel of the input image at the $i^{th}$ row and $j^{th}$ column.

$G_v(i, j)$ is the result of the first convolution.

$G_h(i, j)$ is the result of the second convolution

$y(i, j)$ is the output of the Sobel filtered image at the $i$th row and $j$th column.

The calculations are performed for all $2 \leq i \leq N - 1$ and $2 \leq j \leq M - 1$, where $N$ is the number of rows and $M$ is the number of columns.

Figures A1.2-A1.5 show an example of edge detection using Sobel filter. Figure A1.2 is an image of a square having horizontal and vertical edges. Figure A1.3 shows the edge detection of the image using only the vertical kernel. Figure A1.4 shows the edge detection of the image using only the horizontal kernel. Figure A1.5 shows the edge detection of the image using both - vertical and horizontal - kernels.

Different edge detection algorithms were better under different circumstances. Hence, the algorithm should be judiciously chosen depending upon the requirement. Figure A1.6 shows the results of edge detection using different algorithms on a natural image.
### Appendix I

**PICTURE**

| 0 0 0 0 0 | 160 160 160 160 160 |
| 0 0 0 0 0 | 160 160 160 160 160 |
| 0 0 0 0 0 | 160 160 160 160 160 |
| 0 0 0 0 0 | 160 160 160 160 160 |
| 0 0 0 0 0 | 160 160 160 160 160 |
| 240 240 240 240 240 | 80 80 80 80 80 |
| 240 240 240 240 240 | 80 80 80 80 80 |
| 240 240 240 240 240 | 80 80 80 80 80 |
| 240 240 240 240 240 | 80 80 80 80 80 |
| 240 240 240 240 240 | 80 80 80 80 80 |

**FIGURE A1.2.** Pixel values for an image.

| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 80 | 80 0 0 0 |
| 0 0 0 0 0 80 | 80 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |

**VERTICAL EDGE (G_v)**

| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 80 | 80 0 0 0 |
| 0 0 0 0 0 80 | 80 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |
| 0 0 0 0 0 160 | 160 0 0 0 |

**FIGURE A1.3.** Pixel values for an image formed by edge detection using vertical kernel of Sobel filter.
### HORIZONTAL EDGE (Gₜ)

|   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
|   |   | 0 | 0 | 0 | 0 | 0 | 0 |
|   |   | 0 | 0 | 0 | 0 | 0 | 0 |
|   |   | 0 | 0 | 0 | 0 | 0 | 0 |
| 240| 240| 240| 160|   |   |   |   |
| 240| 240| 240| 160|   |   |   |   |
|   | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|   | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|   | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**FIGURE A1.4.** Pixel values for an image formed by edge detection using horizontal kernel of Sobel filter.

### COMBINED

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td>240</td>
<td>240</td>
<td>240</td>
<td>179</td>
<td>80</td>
</tr>
<tr>
<td>240</td>
<td>240</td>
<td>240</td>
<td>179</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>160</td>
</tr>
</tbody>
</table>

**FIGURE A1.5.** Pixel values for an image formed by edge detection using combination of horizontal and vertical kernel of Sobel filter.
FIGURE A1.6. Edge detection of a natural image using various edge detection algorithms.

The images (b) to (e) have been turned into negative and enhanced using the “AutoBalance” feature of Microsoft Photo Editor.
APPENDIX 2

VQEG SEQUENCES
APPENDIX 2

VQEG SEQUENCES

Video Quality Experts Group (VQEG) has provided twenty standardised test sequences. These test sequences were initially used for the development of the equations for the controller. These test sequences are available in the YUV422 format. In the first stage these sequences were converted to the .ppm files according to the procedure recommended by the VQEG. The collection of .ppm files were then encoded using the MPEG4 codec and used for the development of our controller.

Figure A2.1 shows the best curve fit of the experimental values as discussed in the Chapter 3. The best fit curves were all of the general form:

\[
\text{Bitrate} = A + B \times (\text{Frame Rate}) + C \times (\text{Resolution}) + D \times (\text{Resolution}^2) + E \times (\text{Frame Rate} \times \text{Resolution}) + F \times (\text{Frame Rate} \times \text{Resolution}^2) \quad (A2.1)
\]

Figure A2.2 shows the correlations obtained by plotting the coefficients against the SI and TI. The figure clearly shows wide distribution of the values. Hence, a close match between the calculated and the experimental results cannot be expected.

Figure A2.3 shows the experimental values, best-fit curve and calculated curve using the generalised equation for some VQEG sequences. As expected, the calculated curve does not match the experimental best-fit curve closely. This is due to the wide variations in the test sequences.

Since the method does not work well for a very wide range of sequences, the need arose for the recording of sequences with a more restricted scenario of e-learning.
FIGURE A2.1. Best fit curve for the VQEG sequences. (Cont.)
FIGURE A2.1. Best fit curve for the VQEG sequences. (Cont.)
FIGURE A2.1. Best fit curve for the VQEG sequences. (Cont.)
FIGURE A2.1. Best fit curve for the VQEG sequences. (Cont.)
FIGURE A2.1. Best fit curve for the VQEG sequences. (Cont.)
FIGURE A2.1. Best fit curve for the VQEG sequences. (Cont.)
FIGURE A2.2. Best fit curve for coefficients for VQEG sequences.
FIGURE A2.3. Experimental values, best fit curve and calculated curve using the generalised equation for some VQEG sequences. (Cont.)
APPENDIX 3

Dynamically Adapted Streaming of Video for a Real Time Multimedia Application

IEEE/IARIA Joint ICAS’05 and ICNS’05

International Conference on Autonomic and Autonomous Systems

International Conference on Networking and Services

October 23-28, 2005

Papeete, Tahiti, French Polynesia
Dynamically Adapted Streaming of Video for a Real-Time Multimedia Application

A. Thakur\textsuperscript{1}, B. Lines\textsuperscript{1} and P. Reynolds\textsuperscript{2}

\textit{1. Network Research Group, University of Plymouth, Plymouth, United Kingdom}
\texttt{aruna@jack.see.plymouth.ac.uk; b.lines@plymouth.ac.uk}

\textit{2. Orange SA, United Kingdom}
\texttt{paul.reynolds@orange.co.uk}

\textbf{Abstract}

In most of the synchronous e-learning tools the video of instructor's image is sent using only head-and-shoulder's view. Students find this video presentation rather monotonous and instructors find it restrictive. In this paper we discuss the cinematic principles for automatically capturing a more engaging lecture presentational view. Then we describe our proposed virtual class room set-up for recording the video. Finally we discuss our scheme to stream video sequence having optimum perceptual quality by dynamically controlling frame rate and resolution, taking into considerations the type of video sequence and the required bit rate.

\textbf{1. Introduction}

In recent years the distance learning or e-learning is becoming popular among the students who are located in geographically distant places and who cannot attend classes due to distance or for other reasons. It is especially suited to people who are working and are in need of the further development of their skills but cannot take out time to attend regular classes. The e-learning lectures can be classified into two different categories based upon the interaction between the instructor and students, these are \textit{synchronous} and \textit{asynchronous}. In synchronous e-learning the instruction involves real-time delivery of audio, video, and other media. In asynchronous e-learning educational materials are stored in server and students retrieve them at a time convenient to themselves.

In spite of its growing popularity, the use of video in synchronous e-learning suffers from a number of problems. Some of these problems are:

1. In a real classroom the body language of instructor and the students, like facial expressions, eye gaze and gestures play an important part in nonverbal communication and they direct the attention of students to the relevant area. While in the e-learning videos the students most often only see a head-and-shoulder presentation of their instructor and often with poor quality video. Hence, the instructor is handicapped because he is forced to present it while seated in front of his camera making the class not only very boring but also fails to convey any non-verbal clues. As a result instructors are unable to make the class attractive and guide the students, and students often lose track of the subject.

2. The limited bandwidth available for the transmission of video forces the transmission of video in poor quality. As a result students have to expend extra effort in understanding and deciphering the information.

These deficiencies can result in none or at best limited interaction among participants resulting in boredom and disengagement in students. To overcome these problems two important issues need to be resolved:

(a) Presentation issues. The instruction video should be prepared keeping in mind the principles of cinematography so as to excite the interest of students.

(b) Technical issues. This involves the capturing and transmission of video and audio in a quality comparable to TV quality.

Over the years the art of movie making has evolved so as to make film an effective media of communication. These techniques have also been adopted with certain modifications for "small screen" presentations like TV. Today, even simple events like group discussions, newscast, speeches, etc. can be presented in an interesting and effective way [1].

Unfortunately, most often the instructor just sits in front of a monitor with a camera placed over it and delivers the whole lecture without any movement. Unfortunately, this mere capturing of talking-head-and-shoulder video does not make an interesting
video. What is more interesting than this is a "real lecture" video, where the lecturer has the freedom to move around, use electronic whiteboard and show real models as he / she would do in a real class room environment.

Some of these principles of film making can be adopted for making engaging lecture videos. One of the important cinematographic principles is to use variety of shots from different angles and views. A long duration shot taken from the same angle and view is boring to the audience. Hence, shots of the same object should be taken from different locations and using different zoom. Also each shot should not be of less than 6 seconds and not more than 20 seconds.

2. Aim and Scope of Our Work

While it is possible to produce an exciting video for instructional purposes employing a number of skilled production crew in a studio, it is not economically viable for most colleges on daily basis when they want to deliver lower-cost education via e-learning. Hence, there is a need for the production of cost-effective video following the basic principles of filmmaking. While it may not be possible to automate the "art" of filmmaking; within a limited framework it should be possible to video the classroom lectures in a more interesting manner.

In our project we are developing a system for automating the production and transmission of real time video for synchronous e-learning in an interesting and cost-effective manner. While developing this concept we have defined certain criteria, they are:

1. The hardware for the set-up should be as low budget as possible. Since the idea is to develop a virtual classroom for small colleges having limited resources, we should use inexpensive equipment that is readily available off the shelf.

2. The networking should be done using relatively low bandwidth access technologies.

3. Lecture style should not be made to change too radically. Most of the lecturers have their own style of lecture delivery which they develop over the years. They are not comfortable in changing the style [1]. However in certain circumstances a modification may be necessary.

4. The hardware should be unobtrusive and placed out of the way of instructor and students view [1].

5. The system does not intend to produce "National Geographic quality" instructional video but only to produce video that is more interesting than just a 'talking-head' video.

In this paper we have presented the concepts of our ongoing research work. It has not been completely implemented and we are aware that modifications may be required as we progress with its development.

3. Set-up for a Virtual Class Room

In this section we will only describe the proposed set-up for recording the video of lectures. In order to take shots from different angles and views, the challenging task is the placement and automatic switching of digital video cameras [2]. In our system the video shots can be taken using three different cameras – one fixed and two tracking PTZ (Pan/Tilt/Zoom) cameras. Various techniques have been proposed to switch between the cameras.

Camera 1 is a fixed camera covering the panoramic view that shows the entire lecture dais from which the instructor presents his lectures.

Camera 2 is a tracking PTZ camera to take close-up shot of the instructor’s head and shoulder.

Camera 3 is a tracking PTZ camera to take mid-shot so as to show the instructor and the region of interest next to the instructor, for example the white board, model, etc.

The use of these three cameras for giving different angles / area of view may give the students a sense of presence in real life and create an illusion that the video has been edited by hand [2]. This makes the output video more interesting than that produced by using a single camera to capture head-and-shoulder video.

4. Optimum Transmission of Video

In the proposed scheme, the fluctuation in the quality of video will be mostly due to: (a) changes in video scenes, and (b) random variations in the network conditions. Hence, there is a need for a dynamic adaptive procedure so as to provide consistently good quality video in real-time.

Considerable amount of work has been done on adaptation of video to suit the fluctuations in network. Relatively less work has been done on adaptation of video based both on the network
BEST COPY

AVAILABLE

TEXT IN ORIGINAL IS CLOSE TO THE EDGE OF THE PAGE
conditions and perceptual quality of video. In our previously reported work we discussed the results of the use of a dynamic controller which would predict the complexity (number of macroblocks per frame) of incoming scene and adjust the number of layers and distribution of frames among different layers to be sent so as to maximize the perceptual quality of video [3-5]. We have extended the concept of a controller proposed earlier [3-5] to dynamically adapt to the variations by adjusting the frame rate and spatial resolution. Similar work has been done by other researchers [6-7].

Bit rate of the video sequence is dependent upon the size of frame, frame rate and resolution. Experiments have shown that for a subject to recognize emotions on the face the image should have minimum visual angle of 6 degrees and ideal of 14 degrees. For a student sitting about half a meter from his monitor this approximately corresponds to an image size of 50-125 cm (200-500 pixels) wide, hence we should fix a size in this range [8].

5. Architecture of the System

In our proposed architecture (Figure 1), the video signals from cameras are grabbed and encoded using an open source-code MPEG4 software on Linux platform. The encoded video is then multicast via RTP/UDP/IP stack to establish on-line interactions between server (instructor) and clients (students). A stream from encoder is sent to a feed-forward controller to:

(a) Decide the required bit rate using the information from feedback merger that collects RTCP feedbacks.

(b) Dynamically characterize the nature of video sequence (Temporal Information and Spatial Information).

(c) Use linear prediction for predicting the complexity of the incoming movie sequence.

(d) Decide the optimum way of configuration (frame rate and resolution) of the codec (Figure 2) for the incoming video sequence.

(e) Send the optimized values of configuration parameters to the encoder to configure the codec for coming scene.

Figure 2 schematically shows the flow diagram of the working of the controller.

6. Development of the Controller

Video Quality Experts Group (VQEG) has provided twenty standardized test sequences. These test sequences were used for the development of the equations for the controller. These test sequences are available in the YUV422 format. We have used these video sequences for our work. In the first stage these sequences were converted to the .ppm files according to the procedure recommended by the VQEG. The collection of .ppm files were then encoded using the MPEG4 codec and used for the development of our controller.

6.1. Feedback from Receivers

All the receivers will send RTCP feedbacks indicating the availability of bandwidth. A feedback merger will then combine this data and send the status of the relevant network to the sender. The sender will then use this information to dynamically control the quality of video transmission.
6.2. Characterisation of Video Sequence.

The video sequence are characterized using their Spatial Information (SI) and Temporal Information (TI) as recommended by ITU-T P910 (09/99). The video sequence from the cameras are taken and characterized for SI and TI. The only divergence from the recommendations by ITU is that while ITU recommends taking the highest values of SI and TI of the whole sequence for reporting we keep the individual values of the frames for further calculations by the controller. The SI is based on the Sobel filter and the TI is based on the motion difference feature.

During the calculation ITU recommends that the calculations be performed on a sub-image of the video frame to avoid unwanted edge effects. At the frame rate of 25 fps each frame gets less than 0.04 s for the characterization and optimization. An ordinary PC (2 MB cache, 3 GHz clock speed) requires much more time than that for complete calculation. Hence, to save on the computation time we have to use a sub-image. Comparison between the results obtained for the whole image and the subsections show that up to 50% window size it is possible to obtain an acceptably good match of the values of SI and TI for the whole frame and part of the frame. Further saving on the calculation time may be done by efficient programming and use of faster computer.

The selection of sub-image is done judiciously to allow fast computation and at the same time be representative of the complete frame. For example, in the case of panorama shot most of the frame is occupied by non moving inanimate objects like blackboard, table, etc. There is only a small region where the instructor can move. Hence, only that region is selected for the calculation. This allows the calculation to complete in the available time.

6.3. Prediction of the Quality of Incoming Video Frame

The controller has to predict the quality of incoming frame and then optimize the codec configuration. The SI and TI values of six frames are calculated and then using the concept of linear prediction the expected values of the seventh frame is calculated (Figure 3). The six frames have been chosen for calculation by conducting experiments were the number of frames used for linear prediction was varied from 2 to 15. The use of 6 frames for predicting the seventh was found to be the optimum. Figure 4 shows the SI and TI values of the Sequence 9 and the predicted values that are used for controlling. The figure shows a reasonably good match between the actual and predicted SI and TI values. While the predicted values are being used by the controller, the actual values are stored and used to calculate the next predicted values.

6.4. Development of the Equations for Controller

6.4.1. Relationship between Bitrate – Frame Rate – Resolution

The MPEG4 codec allows encoding at different frame rates and frame sizes. In order to vary the spatial resolution the video sequence was first sub-

![Flow diagram of the prediction of SI and TI of incoming video sequence given the SI and TI of ongoing sequence.](image)

**Figure 3.** Flow diagram of the prediction of SI and TI of incoming video sequence given the SI and TI of ongoing sequence.

![Complexity (SI and TI) of the Sequence 9 and the predicted complexity.](image)

**Figure 4.** Complexity (SI and TI) of the Sequence 9 and the predicted complexity.
sampled to the lower size, while maintaining the aspect ratio, and then up-sampled to the original size. For example, in order to get 50% resolution, the original image having the size 720 pixels x 576 pixels was first sub-sampled to 360 pixel x 288 pixel and then up-sampled back to 720 pixel x 576 pixel. Thus, by varying the frame rate, from 5 to 25 fps in the steps of 5 fps and by varying from 10 to 100% in the steps of 10% a total of 50 different quality encoded sequences of each standard sequence were obtained. The average bitrate of each of these encoded sequence was then taken. Figure 5 shows the plot of Bitrate – Frame Rate – Resolution for the Sequence 9 (Rugby).

In the second step, best fit curve of Bitrate – Frame Rate – Resolution was plotted for the sequences. The best fit curve is of the form:

\[ \text{Bit Rate} = A + B \times \text{(Frame Rate)} + C \times \text{(Resolution)} + D \times \text{(Resolution)}^2 + E \times \text{(Frame Rate} \times \text{Resolution}) + F \times \text{(Frame Rate} \times \text{Resolution})^2 \]

Figure 6 shows an example of the plot.

6.4.2. Relationship between Bitrate – Spatial Information – Temporal Information

Having got all the values of the coefficients for the video sequences, a common equation was developed linking Bitrate – Frame rate – Resolution wherein the coefficients A-F are expressed in terms of the Spatial Information and Temporal Information. It is reasonable to expect that the coefficients of terms including frame rate are functions of the temporal information and the coefficients of the terms including resolution are functions of the spatial information. Hence,

A = f_1(SI, TI); B = f_2(TI); C = f_3(SI);
D = f_4(SI), E = f_5(SI, TI); F = f_6(SI, TI).

We plotted the values of the coefficients against the SI and TI values and obtained the correlations by curve fitting.

Substituting the values of A to F by their respective correlations, a general equation correlating Bitrate – Temporal Information – Spatial Information – Frame rate – Resolution was developed. Figure 7 shows an example of this calculation for the Sequence 9. The figure shows a good match between the actual bitrate and predicted bit rate at lower bit rates. Hence, for the purposes of video transmission for e-learning video it should suffice.

6.4.3. Relationship between Perceptual Quality – Frame rate – Resolution

Extensive work has been done by Cranley et. al. [7] to develop correlation between Perceptual Quality – Frame rate – Resolution. They have developed a general relationship for the Optimum Adaptation Trajectory which defines the correlation between the Frame rate and Resolution which gives the optimum perceptual quality for the given bitrate. According to them:

\[ \text{Resolution} = 21 \ln (\text{Frame Rate}) + 34 \]

We have used this equation for the development of our controller.

6.5. Working of the Controller

Figure 8 shows an example of the decision making process of the controller. In the first step the controller characterizes the ongoing video sequence and then predicts the complexity of the incoming sequence. Based on the predicted SI and TI values it
is possible to calculate the bandwidth requirement for different frame rates and resolutions. The controller then accepts the value of available bandwidth from the feedback merger. Suppose the feedback merger tells the availability of 150 kbps then it is possible to transmit video in 25 different ways, for example (frame rate, resolution) = (7, 100%), (10, 92%), (15, 88%), (20, 82%) and (25, 80%). Out of these 25 ways, the best perceptual quality is obtained by one combination. In the next step that condition is found out using the relationship between Quality - Frame rate - Resolution, which comes out to be (14, 89%). The controller then sends these values to the codec. In this way the controller constantly changes the settings of the codec.

7. Conclusions and Future Work

In this paper we have presented:
- The problems faced in streaming of video in a synchronous e-learning tool.
- Some of the cinematographic principles that may be applied to improve the quality of video.
- Our proposed virtual classroom set-up for transmitting video using a multiple camera system.
- Our scheme for streaming of perceptually optimized and dynamically adapted video over Internet.

In future we will carry out the following:
- In the present work standard video sequences have been used to develop the system, in future the “real” video sequences of a classroom lecture shall be transmitted to the receivers.
- The performance shall be tested objectively and subjectively with and without controller.
- The controller shall be tested in the condition of varying bandwidth.

8. References

APPENDIX 4

UTILIZATION OF THE CONCEPTS OF CHANGE-BLINDNESS FOR EFFICIENT TRANSMISSION OF VIDEO FOR SYNCHRONOUS E-LEARNING APPLICATIONS

WSEAS Transactions on Communications

Issue 11, Volume 4

November 2005

pp. 1176-1186
Utilisation of the Concepts of Change-Blindness for Efficient Transmission of Video for Synchronous E-Learning Applications

ARUNA THAKUR and BENN LINES
Network Research Group
University of Plymouth
Plymouth
UNITED KINGDOM
aruna@jack.see.plymouth.ac.uk; b.lines@plymouth.ac.uk
PAUL REYNOLDS
Orange SA,
UNITED KINGDOM
paul.reynolds@orange.co.uk

Abstract: - E-learning is becoming increasingly popular among the students who are unable to attend regular classes. Unfortunately, in the synchronous e-learning the quality of video is substandard and significant work needs to be done to improve it. This improvement can be brought about either by improving the hardware, like increasing the computer processing power and increasing the bandwidth of network, or by modifying the software, like changing the codec and transmission protocols. The understanding of human visual system can provide a sound basis for improving the quality of video in a synchronous e-learning tool. One of the psycho-physical features of human perceptual system is the “Change Blindness”, which is the inability to note large changes in two similar photos or video sequences. The understanding of change blindness can help in improving the style of presentation and the transmission of video. This paper describes (a) cinematic principles in recording lectures; (b) the concept of change blindness; (c) the ways its understanding can help in making the presentation of video more engrossing; (d) the guidelines it can provide to remove redundant data so as to improve the transmission of video; and (e) proposed scheme to stream video sequence having optimum perceptual quality by dynamically controlling frame rate and resolution.

Key-Words: - change blindness, cinematic principles, background, foreground, separation, perceptual quality, video, transmission

1 Introduction
The Instructional Technology Council defines distance education as “the process of extending learning, or delivering instructional resource-sharing opportunities, to locations away from a classroom, building or site, to another classroom, building or site by using video, audio, computer, multimedia communications, or some combination of these with other traditional delivery methods” [1]. It is a cost-effective way of providing education to students in their own home or office. It is especially suited to students who are located in geographically distant places and who cannot attend classes due to distance or for other reasons. It is becoming popular among the people who are working and are in need of the further development of their skills but cannot take out time to attend regular classes.

The e-learning lectures can be classified into two different categories based upon the interaction between the instructor and students, these are synchronous and asynchronous. In synchronous e-learning the instruction involves real-time delivery of audio, video, and other media. In asynchronous e-learning educational materials are stored in server and students retrieve them at a time convenient to themselves; and the interaction is thus delayed between the instructor and students and involves e-mails, forums, etc.

Due to relative ease in conducting the course work, asynchronous methods have become quite common, but with the improvements in the performance of videoconferencing tools, more and more colleges are experimenting with synchronous methods. In future, some hybrid methods, involving both synchronous and asynchronous methods, may become more common. It is quite possible that virtual classroom, in which the students who can attend will attend the class in person, while those who cannot, will take part by remote connection, could become very popular [2].
2. Some Problems Experienced During the Use of Video in Synchronous e-Learning

In spite of its growing popularity, the use of video in synchronous e-learning suffers from a number of problems.

(a) The courses are ill planned and ill presented since many colleges cannot afford state-of-the-art equipment and the teachers are not properly trained.

(b) Students only see a head-and-shoulder presentation of their instructor and often with poor quality video; so even if the course material is well presented the instructor is handicapped because he is forced to present it while seated in front of his camera.

(c) In a real classroom the body language of instructor and the students, like facial expressions, eye gaze and gestures play an important part in nonverbal communication. Often video fails to convey these clues, as a result instructors are unable to guide the students, and students are unable to give clear feedback.

(d) Even if the instructor does manage to involve students and get them to carry out their tasks satisfactorily, the poor video quality forces the students to expend extra effort in understanding and deciphering the information. This leads to conscious/unconscious discomfort or distress. This distress manifests itself in form of psycho-physiological disorders like mental stress, higher heart beat rate and overall tiredness. This is damaging to the subjects in long run [3].

These deficiencies can result in no or at best limited interaction among participants resulting in boredom and disengagement in students. To overcome these problems two important issues need to be resolved:

- **Presentation issues.** The instruction video should be prepared keeping in mind the principles of cinematography so as to excite the interest of students.
- **Technical issues.** This involves capturing and transmission of video and audio in a quality comparable to TV quality.

Over the years the art of movie making has evolved so as to make film an effective media of communication. These techniques have also been adopted with certain modifications for “small screen” presentations like TV. Today, even simple events like group discussions, newscast, speeches, etc. can be presented in an interesting and effective way; but creating such an interesting and effective video is an art that requires considerable talent and training [4].

Unlike the real classroom, where the students take part in the lectures, in e-learning the control is left primarily in the hands of the instructor. Surveys have shown that students interact with instructor and themselves far less in e-learning courses than in real classes [5]. Hence, it becomes all the more important for instructor to efficiently present the lecture material and guide the students through his presentation without putting too much burden of comprehension on the students.

Considerable amount of work has been done for improving the quality of video transmission. In the hardware side the processing power of the computers has increased manifolds and so has the availability of bandwidth. In the software side, significant amount of work has been done in improving the codecs and transmission protocols. In spite of the tremendous improvements in the last couple of decades, a lot of work needs to be done to make the e-learning videos acceptable to the students.

3 Cinematic Principles in Recording Lectures

Students attending a classroom lecture experience a rich visual experience that is simultaneous, detailed and coherent. That is, if we take a panoramic view of the class we see a number of events at the same time, in high resolution (or high density of information) and all the sections of the image are combined in a proper jigsaw puzzle. But is it really so? Current research in human visual system has shown that this is not the case. The human eye can only see one section of the scene at a time – either in detail or in a coherent (stable) way. We cannot see more than one object in detail and in a coherent way at the same time; and also we can focus attention only on one event at a time [6].

For many years film-makers have informally applied the knowledge of these limitations of human perception to make “continuous” and engaging movies. People watching those films do not complain about the limitations of the perception, on the contrary they find the presentation much more appealing than the real life. If that is the case, then the reason for dull e-learning video lectures is not the limitation due to the use of video for teaching but the actual limitation arises due to the style of presentation. Hence, efforts should be directed towards making video presentation according to the principles by which humans perceive the real world.
By applying the knowledge of strength and weaknesses of human perception we can work towards making the lectures more exciting and captivating.

Unfortunately, not many instructors are trained to be good presenters and video-graphers. Most often the instructor just sits in front of a monitor with a camera placed over it and delivers the whole lecture without any movement. Unfortunately this mere capturing of talking-head-and-shoulder video does not make an interesting video. What is more interesting than this is a “real lecture” video, where the lecturer has the freedom to move around, use electronic whiteboard and show real models as he/she would do in a real class room environment.

Some of these principles of film making can be adopted for making engaging lecture videos, they are:

(a) **Use variety of shots from different angles and views.** A long duration shot taken from the same angle and view is boring to the audience. Hence, shots of the same object should be taken from different locations and using different zoom. Also each shot should not be of less than 6 seconds and not more than 20 seconds.

(b) **Follow a story line.** The role of an instructor is to convey the material in an orderly manner. In a real classroom if this is not done then the instructor can be interrupted and any confusion addressed; on the other hand, such possibilities are far less in the case of e-learning. Hence, the sequence of presentation should be well thought out right at the beginning.

(c) **Intelligent use of the tools to present the course material.** In a real class, students can move around to get proper view of the demonstrations, but the students of e-learning have to rely on the instructor to show them the demonstrations clearly on the screen. Hence, it becomes the responsibility of the instructor to take shots from the appropriate angles and use zoom techniques so to make the object of demonstration clear to the students.

(d) **Continuity of the scene.** While it is important that the continuity of the scene be maintained, it has been observed that small continuity mistakes are often overlooked by the audiences.

### 4 Change Blindness

During the making of a movie different shots are taken at different times and settings. Hence, in spite of all the efforts some continuity mistakes do creep in. In fact there are web pages dedicated to listing these mistakes in popular movies. One of the sites lists the best three continuity mistakes as [7]:

(a) **Commando:** The yellow Porsche is totally wrecked on the left side, until Arnold drives it away, and it is fine.

(b) **Spider-Man:** In the scene where Mary Jane is being mugged by four men, Spider-Man throws two of the men into two windows behind Mary Jane. Then the camera goes back to Spider-Man beating up the other two guys. When the camera goes back to Mary Jane the two windows are intact.

(c) **Terminator 3:** Rise of the Machines: In the scene where John and Catherine are in the hangar at the runway, the Cessna’s tail number is N3035C. When the plane is shown in the air, the number is N9373F. When they land, the tail number has changed back to N3035C.

These are quite glaring mistakes but how many viewers have actually noticed them?

For long researchers have noted that people are quite poor at noticing such changes, but in recent years this subject has renewed considerable interest among researchers and has been researched as “change-blindness”.

According to Simons and Rensink [8], *Change blindness is the striking failure to see large changes that normally would be noticed easily. It is caused when the change is separated by a disturbance. At times even a large and repeatedly made change can go unnoticed for a long time. The disturbance between the scenes may be natural, like the eye movement (saccades) and eye blinks or induced artificially by image flicker, brief “splats” that do not actually cover the region of change. In the case of a movie sequence the disturbance can be caused by a sudden or gradual change in scene. Once the difference has been spotted the change, it becomes very obvious and eye catching [6,8,9].*

The research carried out during the past decade suggests that there are a number of important factors that can cause change blindness.

(a) **Attention:** Change blindness results when the change is unable to attract attention towards itself. This happens more so when the attention is not paid during observation; but in spite of the attention change may often go unnoticed until unless the attention is directed towards the change, which can be aided by a number of ways like using strong or unusual colour, high contrast or verbal clues [8].

(b) **Expectation:** Change blindness is strong when the change is unexpected; this is because the attention is not directed towards the region when the change occurs [8].
Appendix 4 WSEAS Transactions on Communications, pp. 1176-1186, 2005

(c) Background or foreground: Changes to the central item are detected more readily than the changes to the background even if the changes are of equal salience. This is because in general the central item receives more attention as compared to the background [9,10]

(d) Number of changes: Humans have the capability to catch only one change at a time; if more than one change occurs then the change will be noticed one after the other and not at the same time.

(e) Type of Change: It has been shown that the changes to location is more difficult to detect than the changes to the identity, where identity is defined as the objects features other than its position [11,12].

(f) Grouping of objects: It has been known for a long time that the elements are grouped together to form units. It is these units and not individual elements that are further processed for recognition, search, etc. As a result the individual elements lose their own identity [11]. Hence, the memory for the unit is quite good but memory for the individual elements that comprise the unit is poorer. Since the processing is done at the unit level and not at the element level, it has been observed that it is easier to detect changes in visual stimuli that are strongly grouped [11,12].

5.1 Presentation

In a lecture, instructor has to continuously work hard to hold the interest of students. This issue becomes even more important in case of e-learning where the options available to instructor are limited. The knowledge of change-blindness can help the instructor in designing the lecture so that he/she is able to hold the attention of students and make sure that they do not lose track of the subject. Change blindness can play a positive or a negative role in the students' attention. On one hand, due to small attention span, the students' attention can get momentarily diverted away from the object of interest and then due to change-blindness they may forget the earlier scene, thereby resulting in the loss of context. In such a situation the teacher should try to direct the attention of students towards the object of interest by making gestures and sending video sequence taken at the correct angle and zoom of the appropriate object at appropriate time. On the other hand, it can be used purposefully, as done by the film makers and conjurers for a long time, so as to make a well connected presentation and prevent attention getting focused on objects of no interest. Unfortunately, it is not possible for every instructor to be good film-makers. Hence, there is a need for automating the “art” of film-making and this is possible only within a limited domain. The first important step to make the presentation captivating is to break out of the head-and-shoulder mould and present the lecture in a “real” classroom setting but without any interference. In this condition, multiple camera system can take shots from different angles and using different zooms. A combination of fixed and Pan/Tilt/Zoom (PTZ) cameras can cover the lecture dais quite effectively. Once the shots from different angles have been obtained, the second step is to automatically transmit appropriate shot at the appropriate time so as to keep the attention of students focused on the course material and prevent it from wandering away. The transition between the cameras can be done by a number of ways, like manual use of switches, use of gestures, etc. While changing the scene, the transition should be as smooth as possible.

The presentation can be improved further by using a few camera tricks like:

- Use of a very short different shot between two sequences to help in separating different but similar scenes. The very short duration shot does not register in the brain, but it helps to erase the previous scene from the mind, almost like the flicker paradigm. This will help in focusing the
attention on two different objects without interference.

- Once an object has been demonstrated it should be removed from the stage while the camera is not focused on it. Change blindness will cause "out of sight, out of mind"; and its removal will eliminate an object that could potentially be a cause for wandering of attention.

The concept of change blindness can be used in many such ways to improve the quality of presentation.

5.2 Transmission

In e-learning the students often attend the class sitting at their home and are most often linked with very limited bandwidth of up to 60 kbps. Hence, the challenge is to send an acceptable quality of video within that limitation. Significant amount of work has been done in this direction, but inclusion of the concepts of human perceptual model provides a sound basis for further improvements.

When we look at our surrounding we "feel" as if we see a detailed coherent world, and we try to represent this in our video. In actual life this is not so. At a particular instance, we only see either the gist of the panorama or the detailed view of a small part of it. At no instance of time do we have detailed representation of the whole scene. We can apply the same principles for transmission of video.

The quality of video transmission can be significantly improved by reducing the information of the background and enhancing the information of the foreground. Extensive amount of work is being done in separation of foreground and background. This feature is required for various purposes like, tracking of moving object, face recognition, security surveillance, film making, etc. All these applications have different requirements on the quality of separation. In case of transmission, this feature is used to reduce the load on network and has good potential for the use in the transmission of e-learning video, since the students normally have access to limited bandwidth.

In e-learning the separation can be guided by the knowledge of change blindness. Based on the earlier discussion, it is reasonable to expect that the attention of the students would normally be directed towards:

(a) instructor,
(b) blackboard / whiteboard / projection, and
(c) any model or object of interest.

Hence, the segments of the frame covering these objects should be sent preferentially over other parts of the frame. Object of low interest or ungrouped objects like – pointers, etc need not be given much attention and the slight discrepancies in the position of these objects would go unnoticed.

The transmission of important and unimportant sections can be done in various ways, for example:

- The foreground can be sent at high resolution and the background at low resolution;
- The foreground can be sent more frequently as compared to the background;
- In case of panorama shots, high resolution image of the dais can be sent right at the start of the lecture; later on only the foreground can be updated.

6. Aim and Scope of Our Work

While it is possible to produce an exciting video for instructional purposes employing a number of skilled production crew in a studio, it is not economically viable for most colleges on daily basis when they want to deliver lower-cost education via e-learning. Hence, there is a need for the production of cost-effective video following the basic principles of filmmaking. While it may not be possible to automate the "art" of filmmaking; within a limited framework it should be possible to video the classroom lectures in a more interesting manner.

In our project we are developing a system for automating the production and transmission of real time video for synchronous e-learning in an interesting and cost-effective manner.

While developing this concept we have defined certain criteria, they are:

(a) The hardware for the set-up should be as low budget as possible. Since the idea is to develop a virtual classroom for small colleges having limited resources, we should use inexpensive equipment that is readily available off the shelf.

(b) Lecture style should not be made to change. Most of the lecturers have their own style of lecture delivery which they develop over the years. They are not comfortable in changing the style [4]. However in certain circumstances a modification may be necessary.

(c) The hardware should be unobtrusive and placed out of the way of instructor [4].

(d) The system does not intend to produce "National Geographic quality" instructional video but only to produce video that is more interesting than just a "talking-head" video.

In this paper we have only presented the concepts of our ongoing research work. It has not been completely implemented and we are aware that modifications may be required as we progress with
its development. There are basically three parts of our work:
(a) Use of multiple cameras in order to break the monotony of head-and-shoulder shots and to show the object from the appropriate angle and zoom.
(b) Save on bandwidth by removing the redundant features of the frames.
(c) Transmit video within the available bandwidth in the perceptually optimal way.

6.1 Set-up for a Virtual Class Room
In this section we will only describe the proposed set-up for recording the video of lectures. In order to take shots from different angles and views, the challenging task is the placement and automatic switching of digital video cameras [13]. In our system the video shots can be taken using three different cameras – one fixed and two tracking PTZ (Pan/Tilt/Zoom) cameras (Figure 1); it is easy to add more cameras on the same principle. Various techniques have been proposed to switch between the cameras but we will not consider that here. [14]

- Camera 1 is a fixed camera covering the panoramic view that shows the entire lecture dais from which the instructor presents his lectures.
- Camera 2 is a tracking PTZ (Pan/Tilt/Zoom) camera to take close-up shot of the instructor’s head and shoulder.
- Camera 3 is a tracking PTZ camera to take mid-shot so as to show the instructor and the region of interest next to the instructor, for example the white board, model, etc.

The placement of cameras is an important issue. A reasonably good placement of the Camera 1 is at the back from where it can take clean panoramic shots. The Camera 2 takes close-up shots and has the responsibility to show that the instructor is maintaining eye contact with the students. This can be achieved by “direct gaze technique”, which assumes that a remote viewer cannot distinguish whether the instructor is looking at them or not as long as the instructor looks towards the camera within a certain angle. This angle has been found experimentally to be 2.7 degrees horizontally and 9 degrees vertically [5]. Hence, while speaking as far as possible the instructor should look towards the Camera 2. Camera 3 should be placed at the position that will give the best view of the object of interest.

6.2 Separation of Foreground
At the beginning of the session, these background images will be stored both at the sender and receivers ends. During the lecture, the video will be grabbed and the foreground separated from the frames. The foregrounds so generated will be transmitted to the receivers. At the receiver end the foregrounds will be merged with appropriate backgrounds and shown to the receiver. Due to change blindness the receivers may not be able to catch slight discrepancies arising out of the combination of foreground with a general background.

As discussed earlier there are a number of aspects of background separation. A single algorithm cannot cover all the aspects. What is required is an integrated system to cover all the three cases. We are currently working on the development of this system. Primarily it involves the following:

Case 1 Where there is a relative movement between the foreground and background.

There are two sub-cases:
Case 1A Where the foreground is moving and the background is stationary.
Case 1B Where both foreground and background are moving.

Case 2 Where the region of foreground is fixed and represented by objects like blackboard.

The use of these three cameras for giving different angles / area of view may give the students a sense of presence in real life and create an illusion that the video has been edited by hand [13]. This makes the output video more interesting than that produced by using a single camera to capture head-and-shoulder video.
Figure 2. (a) Background; (b) Panorama view of a class; (c) Separation of foreground from background using Algorithm A; (d) Separation of foreground from background using Algorithm B; (e) Separation of foreground from background using a combination of Algorithm A and B; and (f) Identification of edges using Sobel filter.
Where the foreground is moving and the background is stationary.
In the Case 1A, first the background has to be characterised. As a first step the background is grabbed without any foreground. In order to account for the slight variations among different frames arising due to lighting, camera operation, etc. mean and standard deviation of every pixel is taken for a number of frames. As an example, we grabbed 100 frames of the background (Figure 2a) at 25fps. These frames were converted to PPM format and analysed. The mean RGB intensities of the hundred frames were calculated for each pixel. The Figure 3 shows an example of the distribution of the RGB intensities of a pixel about the mean value. As can be seen, the distribution can be modelled as the standard Gaussian distribution.

Figure 2b shows a typical frame with foreground. This foreground can be separated by removing background from the current frame. A number of algorithms have been proposed to do this task. None of these algorithms in themselves can give acceptable separation. A combination of two or more algorithms will give better results. Currently, we are working on the development of a scheme that would give an acceptable separation.

At the initial stages two simple algorithms were implemented.

Algorithm A
In Algorithm A, when the lecture begins, the ongoing frame is compared pixel by pixel against the median or mean background frame. In case the difference is less than a certain predetermined threshold then the pixel is marked as the background else it is marked as the foreground. Figure 3 shows that if the threshold is kept as ±4 then the probability of removing the background is very high. This simple algorithm suffers from a number of shortcomings. Some of the problems are:

- Illumination changes. These can be gradual or sudden. For example, in a classroom illuminated by sunlight there will be a gradual change in the light conditions of the classroom. In an experiment conducted by us we found that the pixel values of background changed by approximately 33 counts over a period of 15 minutes.
- Motion Changes. The motion changes can be due to slight oscillations of the camera.
- Non-ideal distribution. Figure 2 shows that the experimentally found distribution does not follow the ideal Gaussian distribution, hence, while most of the background will be removed, not all the background pixels will be removed.
- Wrong interpretations. If sections of the foreground has values close to that of the background, it will be wrongly interpreted as the background and removed. For example, parts of the foreground’s hands have been wrongly subtracted.

Various modifications of the simple Gaussian subtraction technique have been proposed. For example, in order to account for the gradual changes in the light condition a running average of the background is taken, i.e., the previous background frames are taken as the background for the current frame. Suitable weighted average of the background images may also be used where the recent frames have higher weight. Other more advanced techniques based on Gaussian distribution have also been proposed.

Algorithm B
There are other techniques based on variation of the Gaussian distribution technique. One of them has been proposed by Pan et. al. [15]. In the simplified form, the fourth order variance, i.e., fourth power of the difference between the pixel value of the current frame and the mean of the background frames, is calculated. If the value is greater than a set threshold then the pixel is taken as foreground else it is taken as background. Figure 2d shows the result of separation of the foreground from background based on this algorithm.
Since even Algorithm B is unable to remove all the background pixels, a combination of the two has been used. Figure 2e shows the result of the combination of the two algorithms. While the performance of the combination is better than the individually two algorithms, it is still far from perfect. Hence, better algorithms are required for this task.

Where the region of foreground is fixed
In this case, the region of foreground is fixed and there may or may not be changes to it. In the first step, the outline of the object is detected using the edge detection algorithms and then the region of foreground is decided. Figure 2f shows the result of edge detection using the standard Sobel filter. Further work is in progress in this direction.

6.3 Optimum Transmission of Video
During the transmission of video the fluctuation in the quality of video will be mostly due to: (a) changes in video scenes, and (b) random variations in the network conditions. Hence, there is a need for a dynamic adaptive procedure so as to provide consistently good quality video in real-time. Considerable amount of work has been done on adaptation of video to suit the fluctuations in network. According to one of the approaches the sender partitions the video into a number of layers and the receiver accepts and combines the layers depending upon their individual bandwidth capacity [16]. Relatively less work has been done on adaptation of video based both on the network conditions and perceptual quality of video. In our previously reported work we discussed the results of the use of a dynamic controller which would predict the complexity (number of macroblocks per frame) of incoming scene and adjust the number of layers and distribution of frames among different layers to be sent so as to maximize the perceptual quality of video [17-19]. Similar work has been done by other researchers [20-21].

Bit rate of the video sequence is dependent upon the size of frame, frame rate and resolution. Experiments have shown that for a subject to recognize emotions on the face the image should have minimum visual angle of 6 degrees and ideally of 14 degrees. For a student sitting about half a meter from his monitor this approximately corresponds to an image size of 50-125 cm (200-500 pixels) wide [5], hence we should fix a size in this range.

We have extended the concept of a controller proposed earlier [17-19] to dynamically adapt to the variations by adjusting the frame rate and spatial resolution [14]. In our architecture (Figure 1), the video signals from cameras are grabbed and encoded using an open source-code MPEG4 codec on Linux platform. The encoded video is then multicast via RTP/UDP/IP stack to establish on-line interactions between server (instructor) and clients (students). A stream from encoder is sent to a feed-forward controller to [14]:

(a) Decide the required bitrate using the information from feedback merger that collects RTCP feedbacks.
(b) Dynamically characterize the nature of video sequence (Temporal Information and Spatial Information).
(c) Use linear prediction for predicting the complexity of the incoming movie sequence.
(d) Calculate the options available (frame rate and resolution) that would satisfy the availability of the bandwidth.
(e) Decide the optimum way of configuration (frame rate and resolution) of the codec for the incoming video sequence based on the correlation between the Bandwidth – Frame rate – Resolution – Perceptual quality [20].
(f) Send the optimized values of configuration parameters to the encoder to configure the codec for coming scene.

Figure 4. An example of the working of the controller. For 150 kbps it is possible to have 25 options. Out of these 25 options there is one condition Frame Rate = 14 fps, Resolution = 89% which will give the best perceptual quality.
7. Conclusions and Future Work

In this paper we have discussed the concept of change blindness and how its understanding can help in improving the presentation and transmission of video used for synchronous e-learning. We have also outlined our approach to implement these concepts in the development of e-learning tool.

In future we will work on:
- Development of an appropriate background subtraction technique,
- Implementation of the concept in a real e-learning tool.
- Subjective testing of the concept.

References:


APPENDIX 5

ENHANCING THE QUALITY OF VIDEO IN A SYNCHRONOUS E-LEARNING TOOL BY TAKING INTO ACCOUNT THE CONCEPTS OF CHANGE-BLINDNESS

5th International WSEAS Conference on Distance Learning and Web Engineering

August 23–25, 2005

Corfu, Greece
Enhancing the Quality of Video in a Synchronous E-Learning Tool by Taking into Account the Concepts of Change-Blindness

ARUNA THAKUR and BENN LINES
Network Research Group
University of Plymouth
Plymouth
UNITED KINGDOM
aruna@jack.see.plymouth.ac.uk; b.lines@plymouth.ac.uk

PAUL REYNOLDS
Orange SA,
UNITED KINGDOM
paul.reynolds@orange.co.uk

Abstract: Change blindness is the inability to note large changes in two similar photos or video sequences. In the last decade research carried out to understand this phenomenon has given a new understanding of human perceptual system. This understanding can help in solving some of the problems associated with the transmission of video in a synchronous e-learning tool. The paper describes (a) cinematic principles in recording lectures; (b) the concept of change blindness; (c) the ways its understanding can help in making the presentation of video more engrossing; (d) the guidelines it can provide to remove redundant data and enhance the quality of important sections of the video frame so as to improve the transmission of video; and (e) proposed scheme to stream video sequence having optimum perceptual quality by dynamically controlling frame rate and resolution, taking into considerations the concepts of change blindness.

Key-Words: change blindness, cinematic principles, background, foreground, separation, perceptual quality, video, transmission

1 Introduction
The distance education is a cost-effective way of providing education to students in their own home or office. It is especially suited to students who are located in geographically distant places and who cannot attend classes due to distance or for other reasons. It is becoming popular among the people who are working and are in need of the further development of their skills but cannot take out time to attend regular classes.

The lectures can be classified into two different categories based upon the interaction between the instructor and students, these are synchronous and asynchronous. In synchronous e-learning the instruction involves real-time delivery of audio, video, and other media. In asynchronous e-learning educational materials are stored in server and students retrieve them at a time convenient to themselves. Due to relative ease in conducting the course work, asynchronous methods have become quite common, but with the improvements in the performance of videoconferencing tools, more and more colleges are experimenting with synchronous methods. [1].

In spite of its growing popularity, the use of video in synchronous e-learning suffers from a number of problems. Firstly, in most of the cases the instructor is forced to sit in front of camera and deliver lecture, hence, the students only see a head-and-shoulder presentation of their instructor. As a result, the instructor is handicapped because he is forced to present his class while seated in front of his camera and he cannot make use of the body language, like facial expressions, eye gaze and gestures to guide the students. Secondly, the poor quality video attracts attention to itself and distracts the students from the material being presented. It also leads to conscious / unconscious discomfort or distress. These deficiencies can result in none or at best limited interaction among participants resulting in boredom and disengagement in students.

To overcome these problems two important issues need to be resolved:

(a) Presentation issues. The instruction video should be prepared keeping in mind the principles of cinematography so as to excite the interest of students.
(b) Technical issues. This involves the capturing and transmission of video and audio in a quality comparable to TV quality.

2 Cinematic Principles in Recording Lectures

Students attending a classroom lecture experience a rich visual experience that is simultaneous, detailed and coherent. That is, if we take a panoramic view of the class we see a number of events at the same time, in high resolution (or high density of information) and all the sections of the image are combined in a proper jigsaw puzzle. But is it really so? Current research in human visual system shows that this is not the case. The human eye can only see one section of the scene at a time - either in detail or in a coherent (stable) way. We cannot see more than one object in detail and in a coherent way at the same time; and also we can focus attention only on one event at a time [1].

For many years film-makers have informally applied the knowledge of these limitations of human perception to make “continuous” and engaging movies. People watching those films do not complain about the limitations of the perception, on the contrary they find the presentation much more appealing than the real life. If that is the case, then the reason for dull e-learning video lectures is not the limitation due to the use of video for teaching, but the actual limitation arises due to the style of presentation. Hence, efforts should be directed towards making video presentation according to the principles by which humans perceive the real world. By applying the knowledge of strength and weaknesses of human perception we can work towards making the lectures more exciting and captivating.

Unfortunately, not many instructors are trained to be good presenters and video-graphers. Most often the instructor just sits in front of a monitor with a camera placed over it and delivers the whole lecture without any movement. Unfortunately this mere capturing of talking-head-and-shoulder video does not make an interesting video. What is more interesting than this is a “real lecture” video, where the lecturer has the freedom to move around, use electronic whiteboard and show real models as he / she would do in a real class room environment.

Some of these principles of film making can be adopted for making engaging lecture videos, they are:

(a) Use variety of shots from different angles and views. A long duration shot taken from the same angle and view is boring to the audience. Hence, shots of the same object should be taken from different locations and using different zoom. Also each shot should not be of less than 6 seconds and not more than 20 seconds.

(b) Follow a story line. The role of an instructor is to convey the material in an orderly manner. In a real classroom if this is not done then the instructor can be interrupted and any confusion addressed; on the other hand, such possibilities are far less in the case of e-learning. Hence, the sequence of presentation should be well thought out right at the beginning.

(c) Intelligent use of the tools to present the course material. In a real class, students can move around to get proper view of the demonstrations, but the students of e-learning have to rely on the instructor to show them the demonstrations clearly on the screen. Hence, it becomes the responsibility of the instructor to take shots from the appropriate angles and use zoom techniques so to make the object of demonstration clear to the students.

(d) The continuity of the scene. While it is important that the continuity of the scene be maintained, it has been observed that small continuity mistakes are often overlooked by the audiences.

3 Change Blindness

During the making of a movie different shots are taken at different times and settings. Hence, in spite of all the efforts some continuity mistakes do creep in. In fact there are web pages dedicated to listing these mistakes in popular movies. One of the sites lists the best three continuity mistakes as [2]:

(a) Commando: The yellow Porsche is totally wrecked on the left side, until Arnold drives it away, and it is fine;

(b) Spider-Man: In the scene where Mary Jane is being mugged by four men, Spider-Man throws two of the men into two windows behind Mary Jane. Then the camera goes back to Spider-Man beating up the other two guys. When the camera goes back to Mary Jane the two windows are intact.

(c) Terminator 3: Rise of the Machines: In the scene where John and Catherine are in the hangar at the runway, the Cessna's tail number is N3035C. When the plane is shown in the air, the number is N9373F. When they land, the tail number has changed back to N3035C.
These are quite glaring mistakes but how many viewers actually noticed them? For long researchers have noted that people are quite poor at noticing such changes, but in recent years this subject has renewed considerable interest among researchers and has been researched as "change-blindness".

According to Simons and Rensink [3], Change blindness is the striking failure to see large changes that normally would be noticed easily. It is caused when the change is separated by a disturbance. At times even a large and repeatedly made change can go unnoticed for a long time. The disturbance between the scenes may be natural, like the eye movement (saccades) and eye blinks or induced artificially by image flicker, brief "splats" that do not actually cover the region of change. In the case of a movie sequence the disturbance can be caused by a sudden or gradual change in scene. Once the difference has been spotted the change, it becomes very obvious and eye catching [1,3,4]. The research carried out during the past decade suggests that there are a number of important factors that can cause change blindness.

(a) Attention: Change blindness results when the change is unable to attract attention towards itself. This happens more so when the attention is not paid during observation; but in spite of the attention change may often go unnoticed until the attention is directed towards the change, which can be aided by a number of ways like using strong or unusual colour, high contrast or verbal clues [3].

(b) Expectation: Change blindness is strong when the change is unexpected; this is because the attention is directed towards the region when the change occurs [3].

(c) Background or foreground: Changes to the central item are detected more readily than the changes to the background even if the changes are of equal salience. This is because in general the central item receives more attention as compared to the background [4,5].

(d) Number of changes: Humans have the capability to catch only one change at a time; if more than one change occurs then the change will be noticed one after the other and not at the same time.

(e) Type of Change: It has been shown that the changes to location are more difficult to detect than the changes to the identity, where identity is defined as the objects features other than its position [6,7].

(f) Grouping of objects: It has been known for a long time that the elements are grouped together to form units. It is these units and not individual elements that are further processed for recognition, search, etc. As a result the individual elements lose their own identity [6]. Hence, the memory for the unit is quite good but memory for the individual elements that comprise the unit is poorer. Since the processing is done at the unit level and not at the element level, it has been observed that it is easier to detect changes in visual stimuli that are strongly grouped [6,7].

A number of models have been suggested to define human perceptual system and explain the existence of change-blindness. The complexity of the task means that none of these models are able to explain all the observations convincingly.

While change blindness has attracted considerable interest among the psychologists, it has gone largely unnoticed by computer scientists. In fact it opens a large unexploited domain of study. In this paper we have tried to explain how the understanding of change blindness can help in designing e-learning courses and the work that we are carrying out to use the understanding of change blindness to develop an efficient system for transmission of video.

4 Lessons for e-Learning

The research on change-blindness raises some interesting prospects for improving the quality of e-learning video both in terms of presentation and transmission.

4.1 Presentation

In a lecture, instructor has to continuously work hard to hold the interest of students. This issue becomes even more important in case of e-learning where the options available to instructor are limited. The knowledge of change-blindness can help the instructor in designing the lecture so that he/she is able to hold the attention of students and make sure that they do not lose track of the subject. Change blindness can play a positive or a negative role in the students' attention. On one hand, due to small attention span, the students' attention can get momentarily diverted away from the object of interest and then due to change-blindness they may forget the earlier scene, thereby resulting in the loss of context. In such a situation the teacher should try to direct the attention of students towards the object of interest by making gestures and sending video sequence taken at the correct angle and zoom of the appropriate object at appropriate time. On the other
hand, it can be used purposefully, as done by the film makers and conjurers for a long time, so as to make a well connected presentation and prevent attention getting focused on objects of no interest. Unfortunately, it is not possible for every instructor to be good film-makers. Hence, there is a need for automating the “art” of film-making and this is possible only within a limited domain. The first important step to make the presentation captivating is to break out of the head-and-shoulder mould and present the lecture in a “real” classroom setting but without any interference. In this condition, multiple camera system can take shots from different angles and using different zooms. A combination of fixed and Pan/Tilt/Zoom (PTZ) cameras can cover the lecture dais quite effectively. Once the shots from different angles have been obtained, the second step is to automatically transmit appropriate shot at the appropriate time so as to keep the attention of students focused on the course material and prevent it from wandering away. The transition between the cameras can be done by a number of ways, like manual use of switches, use of gestures, etc. While changing the scene, the transition should be as smooth as possible.

The presentation can be improved further by using a few camera tricks like:

- Use of a very short different shot between two sequences to help in separating different but similar scenes. The very short duration shot does not register in the brain, but it helps to erase the previous scene from the mind, almost like the flicker paradigm. This will help in focusing the attention on two different objects without interference.

- Once an object has been demonstrated it should be removed from the stage while the camera is not focused on it. Change blindness will cause “out of sight, out of mind”, and its removal will eliminate an object that could potentially be a cause for wandering of attention.

The concept of change blindness can be used in many such ways to improve the quality of presentation.

4.2 Transmission

In e-learning the students often attend the class sitting at their home and are most often linked with very limited bandwidth of up to 60 kbps. Hence, the challenge is to send an acceptable quality of video within that limitation. Significant amount of work has been done in this direction, but inclusion of the concepts of human perceptual model provides a sound basis for further improvements.

When we look at our surrounding we “feel” as if we see a detailed coherent world, and we try to represent this in our video. In actual life this is not so. At a particular instance, we only see either the gist of the panorama or the detailed view of a small part of it. At no instance of time do we have detailed representation of the whole scene. We can apply the same principles for transmission of video.

The quality of video transmission can be significantly improved by reducing the information of the background and enhancing the information of the foreground. Extensive amount of work is being done in separation of foreground and background. This feature is required for various purposes like, tracking of moving object, face recognition, security surveillance, film making, etc. All these applications have different requirements on the quality of separation. In case of transmission, this feature is used to reduce the load on network and has good potential for the use in the transmission of e-learning video, since the students normally have access to limited bandwidth.

In e-learning the separation can be guided by the knowledge of change blindness. Based on the earlier discussion, it is reasonable to expect that the attention of the students would normally be directed towards:

(a) instructor,
(b) blackboard / whiteboard / projection, and
(c) any model or object of interest.

Hence, the segments of the frame covering these objects should be sent preferentially over other parts of the frame. Object of low interest or ungrouped objects like – pointers, etc need not be given much attention and the slight discrepancies in the position of these objects would go unnoticed.

The transmission of important and unimportant sections can be done in various ways, for example:

- The foreground can be sent at high resolution and the background at low resolution;
- The foreground can be sent more frequently as compared to the background;
- In case of panorama shots, high resolution image of the dais can be sent right at the start of the lecture; later on only the foreground can be updated.

5. Aim and Scope of Our Work

In our project we are developing a system for automating the production and transmission of real time video for synchronous e-learning in an interesting and cost-effective manner. In this paper we have only presented the concepts of our ongoing
research work. It has not been completely implemented and we are aware that modifications may be required as we progress with its development. There are basically three parts of our work:

(a) Use of multiple cameras in order to break the monotony of head-and-shoulder shots and to show the object from the appropriate angle and zoom.
(b) Save on bandwidth by removing the redundant features of the frames.
(c) Transmit video within the available bandwidth in the perceptually optimal way.

5.1 Use of Multiple Cameras and Automatic Selection of Shots

For the first part, we have proposed a three camera system; it is easy to add more cameras on the same principle. Camera 1 is a fixed camera covering the panoramic view that shows the entire lecture dais from which the instructor presents his lectures. Camera 2 is a tracking PTZ (Pan/Tilt/Zoom) camera to take close-up shot of the instructor’s head and shoulder. Camera 3 is a tracking PTZ camera to take mid-shot so as to show the instructor and the region of interest next to the instructor, for example the white board, model, etc. [8].

![Figure 1. Block diagram of the architecture.](image)

5.2 Separation of Foreground

As discussed earlier there are a number of aspects of background separation. A single algorithm cannot cover all the aspects. What is required is an integrated system to cover all the three cases. We are currently working on the development of this system. Primarily it involves the following:

(a) Separation of foreground, where there is a relative movement between the foreground and background. This has two sub-cases: (i) Where the foreground is moving and the background is stationary; and (ii) Where both foreground and background are moving. In the first case, first the background is grabbed without any foreground. In order to account for the slight variations among different frames arising due to lighting, camera operation, etc. mean and standard deviation of every pixel is taken for a number of frames. The variations normally follow the standard Gaussian curve. A threshold of variations is then decided. When the lecture begins, the ongoing frame is compared pixel by pixel against the mean background frame, and a pixel whose value is out of the set threshold is taken as the foreground. The second case in more complex and a few algorithms have been developed to separate the two. Further work needs to be done in this case.

(b) Foreground as represented by objects like blackboard. In this case the region of foreground is fixed and there may or may not be changes to it. In the first step, the outline of the object is detected using the edge detection algorithms and then the region of foreground is decided.

At the beginning of the session, these background images will be stored both at the sender and receivers ends. During the lecture, the video will be grabbed and the foreground separated from the frames. The foregrounds so generated will be transmitted to the receivers. At the receiver end the foregrounds will be merged with appropriate backgrounds and shown to the receiver. Due to change blindness the receivers may not be able to catch slight discrepancies arising out of the combination of foreground with a general background.

![Figure 2. (a) Background; (b) Panorama view of a class; (c) Separation of foreground from background; and (d) Identification of edges using Sobel filter.](image)
5.3 Optimum Transmission of Video
During the transmission of video the fluctuation in the quality of video will be mostly due to: (a) changes in video scenes, and (b) random variations in the network conditions. Hence, there is a need for a dynamic adaptive procedure so as to provide consistently good quality video in real-time.

We have developed a controller to dynamically adapt to the variations by adjusting the frame rate and spatial resolution [8]. In our architecture (Figure 1), the video signals from cameras are grabbed and encoded using an open source-code MPEG4 codec on Linux platform. The encoded video is then multicast via RTP/UDP/IP stack to establish on-line interactions between server (instructor) and clients (students). A stream from encoder is sent to a feed-forward controller to [8]:
(a) Decide the required bitrate using the information from feedback merger that collects RTCP feedbacks.
(b) Dynamically characterize the nature of video sequence (Temporal Information and Spatial Information).
(c) Use linear prediction for predicting the complexity of the incoming movie sequence.
(d) Calculate the options available (frame rate and resolution) that would satisfy the availability of the bandwidth.
(e) Decide the optimum way of configuration (frame rate and resolution) of the codec for the incoming video sequence based on the correlation between the Bandwidth – Frame rate – Resolution – Perceptual quality [9].
(f) Send the optimized values of configuration parameters to the encoder to configure the codec for coming scene.

8. Conclusions and Future Work
In this paper we have discussed the concept of change blindness and how its understanding can help in improving the presentation and transmission of video used for synchronous e-learning. We have also outlined our approach to implement these concepts in the development of e-learning tool.

References:
APPENDIX 6

A DYNAMIC CONTROLLER FOR OPTIMAL LAYERING OF VIDEO

ACM Multimedia 2002

10th International Conference on Multimedia

December 1-6, 2002

Juan les Pins, France
A Dynamic Controller for Optimal Layering of Video

Aruna Thakur
Department of Computer Science and Electrical Engineering
Luleå University of Technology
SE – 971 87, Luleå, Sweden
aruna@cdt.luth.se
Advisor: Lenka Carr-Motyckova
Date of beginning May 2001
Expected end of PhD work May 2004

ABSTRACT
Layering techniques for video transmission are being actively considered as a way of satisfying receivers in a heterogeneous environment. A layering scheme based on a combination of network constrains and the perceptual quality of video is an attractive proposition. In this paper a feed-forward post-codec controller has been described for sending a video sequence in an optimally layered way. For optimization the controller considers the perceptual quality of video, load on sender and load on network. The correlations used for the controller have been developed experimentally and deductively. The controller so developed has been shown to enhance the perceptual quality of video at the receiver end in a heterogeneous environment.

Keywords
Layering, perceptual quality, codec, controller.

1. INTRODUCTION
In this ongoing work the main idea is to develop a layering technique for an inexpensive video conferencing tool suitable for use in distance education and healthcare [1]. For this purpose the effect of various layering parameters, for example, the perceptual quality of video and load on network / sender has been experimentally studied. From the results obtained, an architecture has been developed that dynamically calculates the optimal layering technique [2]. It has been observed that the perceptual quality of video at the receiver end is better in the system with controller than the system without controller [3].

2. ARCHITECTURE OF THE SYSTEM
The architecture has been developed for the optimal post-codec temporal layering using Intra H.261 codec, but with modifications it can be made more general and used for other codecs also. In this architecture of a feed-forward system, a controller dynamically characterizes the nature of video sequence, takes input from the receivers regarding the network condition and then divides the video sequence into layers in an optimal way before transmitting it into network. In such a controller the dependent variables are (a) load on sender, (b) load on network, (c) perceptual quality of video; and the independent variables are (i) complexity of video sequence (ii) number of layers, (iii) distribution of frames among different layers, and (iv) frame rate.

In this work the complexity of a video sequence has been defined as the number of Macroblock sent per frame. The frames have been distributed according to two algorithms (a) Algorithm A, where the successive frames are divided equally among different layers, and (b) Algorithm B, where more frames are sent to higher numbered layers so that \( N_L \) (No. of frames in a layer \( L \)) \( \leq 2^{L-1} \).

Figure 1 shows the block diagram for the architecture. The sender sends the video stream to the codec. The encoded stream goes to a demultiplexer, where the header of the RTP packets is read. On the basis of Macroblock address the demultiplexer partitions the stream and sends them to different IP multicast addresses. At the receiver end, the multiplexer combines all the accepted layers and sends them to the decoder.

The complexity calculator in the controller reads the number of Macroblock per frame for a set of frames. Using the concept of linear prediction the expected average number of Macroblock per frame for the next set of frames is calculated. While the predicted value of the complexity is being used by the controller, the actual values are stored and used to calculate the next predicted value. A set of frames has the same number of frames as the number of layers in which the video sequence shall be divided into. Figure 2 shows schematically the working of the complexity calculator. Figure 3 shows complexity (number of Macroblock per frame) versus frame number of a video sequence of approximately 2 minutes. The figure also shows the prediction for the video sequence for the fixed number (3) of frames per set.

In combination with the feedback from the receivers regarding network conditions and the characteristics of the short video sequence, various parameters affecting the layering are optimized. This is done by first finding the various layering options available that satisfy the bandwidth requirement for a receiver, out of the available options the option with best perceptual quality is selected. Out of the various options available to the group of receivers the one that satisfies best is finally selected.

The relationship between the complexity of video, number of layers and distribution schemes has been found empirically using experimental tests [2]. The effect of the parameters like load on...
judgment finer. The subjects were asked to consider the other video sequences were judged in comparison to this uncompressed video clip with a maximum quality of 10, and all scale between 1-10, instead of 1-5. This was done to make the T standard . A slight modification was made so as to have quality was subjectively tested according to the recommendations of ITU-

The video was sent in I to 6 layers. The perceptual quality of video frames were divided into layers. In the present investigation the timestamp of the frame, and then based on the algorithm the subsequently to the demultiplexer. The demultiplexer read the video from a camera / DVD / VCR was sent to the codec and 3-I. Development of the Controller

3. EXPERIMENTAL PROCEDURE

3.1. Development of the Controller

Video from a camera / DVD / VCR was sent to the codec and subsequently to the demultiplexer. The demultiplexer read the timestamp of the frame, and then based on the algorithm the frames were divided into layers. In the present investigation the video was sent in 1 to 6 layers. The perceptual quality of video was subjectively tested according to the recommendations of ITU-T standard. A slight modification was made so as to have quality scale between 1-10, instead of 1-5. This was done to make the judgment finer. The subjects were asked to consider the uncompressed video clip with a maximum quality of 10, and all other video sequences were judged in comparison to this sequence. For each type of video sequence 15 people were asked to give their opinion. Audio was muted in the experiment. The experimental values had standard deviation less than 1.5. The load on sender was taken as the mean of CPU load. The load on network was found by taking the mean of the values of available bandwidth (kbps) used at the receiver end of the conferencing tool. Two sets of frame rates have been tested: 3 fps and 8 fps. The main reason for choosing such low frame rate was that terminals connected to low link capacity commonly work on such low frame rate. In the development stages of the controller [2], different movies with varied range of complexity were used to form an empirical evaluation of the bandwidth, perceptual quality, load on network and performance of the algorithms A and B [2]. The equations so developed were used by the controller for activating the algorithms A and B as required

3.2. Testing of the Controller.

Two different types of configurations were tested. The sender was connected to a router by a link with a capacity of 100 Mbps. A multicast router was a Red Hat Linux (version 7.3 2.96-110, Kernal version 2.4.18-3) PC system in which one or more class-based differentiated services are implemented where the class based queuing (CBQ) was performed. The router was used to limit the bandwidth for base and enhancement layers. A video sequence of 2.5 minutes was sent through the protocol at the receiver end. The movie sequence was so chosen to have sections having low (2 Macroblock per frame) to high (15 Macroblock per frame) complexity. Three different tests were carried out: (a) using only Algorithm A, (b) using only Algorithm B and (c) letting the controller choose between the algorithms A and B.

In the Configuration 1 there was one sender and one receiver. The receiver was connected to the router with a link having variable capacity (20, 40, 60 and 80 kbps). In the Configuration 2 there was only one sender but four receivers. All the receivers were connected in parallel and had link capacities of 20, 40, 60 and 80 kbps. The same video sequence as the previous case was used and the same three tests were carried out.

To judge the perceptual quality as a reference the same movie was sent without any layering. This movie was given a value 0, and the test video sequence were judged in relation to this. The scale varied between +5 (better performance) to −5 (worse performance).

4. RESULTS AND DISCUSSION

4.1. Effect of Number of Layers

Figure 4 shows an example of the effect of layering of video sequence using Algorithms A. The subjects noted the deterioration in the quality with the increase in the number of layers even when the receiver accepted all the fractional layers. The quality also deteriorated with the decrease in the number of fractional layers accepted. Figure 5 shows an example of the effect of number of layers on the load on network. As expected, the load on network increases with increase in number of layers accepted.

4.2. Effect of Frame Distribution

Figures 4 and 6 show the effect of layering on the perceptual quality of video for the algorithm A and B. As is to be expected Algorithm A gives a linear increase in perceptual quality with...
receiver accepted different number of layers. The figure shows that as the complexity increases the perceptual quality falls. The reason may be that the video sequence with low complexity need less bandwidth, hence, they do not suffer much loss of quality. Secondly, the slow rate of changes makes the movement seem smoother to the viewers [2].

4.4. Performance of the Controller
Figure 3 is a representative example of the video sequence sent through the 20 kbs link in Configuration 1. The figure clearly shows the switching between the two algorithms in the controller depending on the complexity at a given instance. Comparison between the two sections of the figure shows that at lower complexity the controller sends the video in one layer using Algorithm A and at higher complexity the controller sends the video in two layers using Algorithm B.

It has been observed [3] that when the video sequence is sent without layering there are random packet drops depending upon the available bandwidth resulting in the deterioration of the video quality. As expected, the lesser the bandwidth available poorer is the quality. As Figure 8 shows, in this case the subjects reported better performance for the video sequence with layering compared to without layering in the case of 20, 40 and 60 kbps links. They did not report any improvement in the performance for 80 kbps link. This may be because the Algorithm B works better at the lower bandwidths. At the 80 kbps link again there are no packet drops, hence, no effect of layering.

5. CONCLUSIONS
From the study carried out it may be concluded that a new architecture for dynamic division of video into layers has been developed. For the division the controller takes into account: (a) complexity of the scene, (b) perceptual quality, (c) load on sender, (d) load on network; and the controller optimises (i) number of layers, frame rate, (ii) distribution of frames among different layers. Use of the controller enhances the perceptual quality of video at the receiver end in a heterogeneous environment.

6. PLANS FOR FUTURE WORK
In the future we want to introduce an online objective test to calculate the perceptual video quality at both the sender and receivers end and use the result to judge the optimum number of layers. We will also use the concept for spatial layering. Similar concept shall be tried using MPEG4.

7. REFERENCES
APPENDIX 7

FEED-FORWARD CONTROLLER FOR LAYERED VIDEO CODING

WWW/Internet 2002

IADIS International Conference

November 13-15, 2002

Lisbon, Portugal
FEED-FORWARD CONTROLLER FOR LAYERED VIDEO CODING

Aruna Thakur, Lenka Motyckova-Carr and Per-Erik Bjork
Department of Computer Science and Electrical Engineering
Luleå University of Technology
SE – 971 87, Luleå, Sweden
aruna@cdt.luth.se, (lenka.phb)@sm.luth.se

ABSTRACT
Layering techniques for video transmission are being actively considered as a way of satisfying receivers in a heterogeneous environment. A layering scheme based on a combination of network requirements and the perceptual quality of video is an attractive proposition. In this paper a feed-forward post-codec controller has been described that sends video sequence in an optimally layered way. For optimization the controller considers the perceptual quality of video, load on sender and load on network. For the controller the correlations have been developed experimentally and deductively.

KEYWORDS
Layering, perceptual quality, codec, controller,

1. INTRODUCTION
In the layered transmission of video, the video is divided into a number of layers [1]. The first layer, the base layer, contains the basic information about the video, and the other layers, the enhancement layers, provide the information for the quality improvement. Depending upon the availability of bandwidth, the receiver can accept one or more layers [3].

The knowledge of proper layering technique is of importance to the developers of Internet - based conferencing tools. In the present network, private subscribers often have rather low availability of bandwidth. Hence, these tools work quite satisfactorily using rather “old” codec like Intra-H.261, which is designed for low bit rate transmission. In order to solve the problems related to temporal synchronization, Intra-H.261 uses temporal compression model called conditional replenishment. Only those blocks that change substantially in successive frames are encoded and sent. Hence, as the changes taking place in a sequence increases the number of macroblocks sent per frame also increases. This puts additional burden on the sender and on the network. The rate of transmission of macroblocks per frame also affects the perceptual quality.

The quality of video can be judged either subjectively or objectively. For subjective test, in the absence of any other standard, the ITU-R and ITU-T, standards for digital TV may be used. There are a number of objective tests. Some of these, like peak signal to noise ratio (PSNR) and root mean square error (RMSE), are simple but do not correlate well with the subjective evaluation. Other more complex matrices are based on the complex spatial and temporal model of human visual system. These models evaluate the video much more accurately but are more difficult to work with.

The main idea of our study is to develop best layering technique for an inexpensive video conferencing tool suitable for general use in distance education and healthcare [5]. For this purpose, the effect of various layering parameters on the perceptual quality [2] of video and load on network / sender has been experimentally studied. From the results, an architecture has been developed that dynamically calculates the optimal layering technique.
2. ARCHITECTURE OF THE SYSTEM

The architecture has been developed for the optimal post-codec temporal layering in Intra H.261 codec, but with modifications it can be made more general and used for other codecs also. In this architecture of a feedforward system a controller dynamically characterizes the nature of video sequence, takes input from the receivers regarding the network condition and then divides the video sequence into layers in an optimal way before transmitting it into network. In such a controller the dependent variables are (a) load on sender, (b) load on network, (c) perceptual quality of video; and the independent variables are (a) complexity of video sequence (b) number of layers, (c) distribution of frames among different layers, (d) frame rate.

In this work the complexity of video sequence has been defined as the number of macroblocks sent per frame. The frames have been distributed according to two algorithms (a) Algorithm A, the successive frames are sent equally on the different layers, and (b) Algorithm B, more frames are sent to higher numbered layers so that \( N_L \) (No. of frames in a layer \( L \)) \( \propto 2^{L-1} \).

Figure 1 shows the block diagram for the architecture. The sender sends the video stream to an Intra-H.261 codec. The encoded stream goes to a demultiplexer, where the header of the RTP packets is read. On the basis of macroblock address the demultiplexer partitions the stream and sends them to different IP multicast addresses. At the receiver end, the multiplexer combines all the accepted layers and sends them to the decoder.

In the complexity calculator the number of macroblocks per frame is read for a set of frames. Using the concept of linear prediction the expected average number of macroblocks per frame for the next set is calculated. This predicted value is used by the controller to calculate the number of layers, etc. While the predicted value is being used by the controller, the actual complexity is stored and used to calculate the next predicted value. A set of frames has the same number of frames as the number of layers in which the video sequence shall be divided into. Figure 1 also shows schematically the working of the complexity calculator. Figure 2 shows complexity (number of macroblocks per frame) versus frame number of a video sequence of approximately 2 minutes. The figure also shows the prediction for the video sequence for the fixed number (4) of frames per set.

In combination with the feedback [4] from the receivers regarding the network conditions and the characteristics of the short video sequence various parameters affecting the layering are optimized using iterative calculations. Having found the optimum distribution scheme the controller directs the demultiplexer to distribute the video sequence into layers. The controller uses various empirical and deductive equations between the independent and dependent variables for the calculations.

In order to have co-ordination between the sender and receiver, before changing the number of layers the sender first sends a message to the receivers regarding the change. The receiver then adjusts its existing position to a nearest bandwidth equivalent position to the new system by using the conversion database present with the receiver. In this work feedback from only one receiver is being considered. In a heterogeneous environment a feedback merger can be employed to take into account the different requirements of the receivers. Since in a videoconference the number of receivers is rather limited, the characteristics of each receiver can be fed to the controller so that it can make adjustments so as to satisfy the group in the best possible way. In the case of many receivers the statistical characteristics can be given to the controller.

3. EXPERIMENTAL PROCEDURE

Video from a camera / DVD player / VCR was sent to the codec and subsequently to the demultiplexer. The demultiplexer read the timestamp of the frame, and then based on the algorithm the frames were divided into layers. In the present investigation, the video was sent in 1 to 6 layers. The perceptual quality of video was subjectively tested according to the recommendations of ITU-R standard. A slight modification was made so as to have quality scale between 1-10, instead of 1-5. This was done to make the judgment finer. The subjects were asked to consider the uncompressed video clip with a maximum quality of 10, and all other video sequence were judged in comparison to this sequence. For each type of video sequence 15 people were asked
Figure 1. Block diagrams of the proposed architecture and complexity calculator. 1-5: complexity switch, A-C: algorithms used for distributing the frames among various layers, items in the box need to be optimized before the controller adjusts the demultiplexer, Fx, F(x+N): No. of macroblocks in frame numbers, x, (x+N) respectively, N: No. of layers in the previous set.

Figure 2: Complexity (number of macroblocks per frame) of a video sequence of approximately two minutes and the complexity used by controller to give their opinion. Audio was muted in the experiment. The experimental values had standard deviation less than 1.5. Figure 3 shows an example of the transmitted video sequence. The load on sender was taken as the mean of CPU load. The load on network was found by taking mean of the values of bandwidth (kbps) used at the receiver end of the conferencing tool. Two sets of frame rates have been tested: 3.33 fps and 7.78 fps.

4. RESULTS AND DISCUSSION

4.1. Effect of Video Complexity

Figure 4 shows the effect of complexity on the perceptual quality of video when the video sequence was sent in 6 layers and the receiver accepted different number of layers. The figure shows that as the complexity increases the perceptual quality falls. The reason may be that the video sequence with low complexity need less bandwidth, hence, they do not suffer much loss of quality. Secondly, the slow rate of changes makes the movement seem smoother to the viewers.

4.2. Effect of Number of Layers

Figure 5 shows an example of the effect of layering of video sequence using Algorithms A. The subjects noted the deterioration in the quality with the increase in the number of layers even when the receiver accepted all the fractional layers. The quality also deteriorated with the decrease in the number of fractional layers accepted. Figure 6 shows an example of the effect of number of layers on the load on network. As expected, the load on network increases with increase in number of layers accepted.
4.3. Effect of Frame Distribution

Figures 5 and 7 show the effect of layering on the perceptual quality of video for algorithms A and B. As is to be expected Algorithm A gives a linear increase in perceptual quality with increase in number of layers, but Algorithm B gives exponential increase in quality with increase in number of layers. An interesting observation is that from the point of view of perceptual quality the most significant improvement in the quality takes place when the last fractional layer is joined. This is more so in the case of algorithm B where the more weight is given to higher layers in terms of bandwidth distribution.

4.4. Effect of Frame Rate

The quality improves with increase in the frame rate, but at the expense of load on network.

Using regression analysis, equations that best fit the experimental data have been developed. Equations have been developed for various dependent and independent variables.

5. CONCLUSIONS

From the study carried out it may be concluded that the new architecture for dynamic division of video into layers has been developed. The controller takes into account: (a) complexity of the scene, (b) perceptual quality, (c) load on sender, (d) load on network; and the controller optimises (a) number of layers, frame rate, (e) distribution of frames among different layers.

6. REFERENCES

APPENDIX 8

PERFORMANCE OF A DYNAMIC CONTROLLER FOR LAYERED STREAMING OF VIDEO IN INTERNET

IEEE/IEE ICT 2003

10th International Conference on Telecommunications

February 23-March 1, 2003

Tahiti, Papeete, French Polynesia
Performance of a Dynamic Controller for Layered Streaming of Video in Internet

Aruna Thakur and Lenka Carr - Motyckova
Department of Computer Science and Electrical Engineering
Luleå University of Technology
SE – 971 87, Luleå, Sweden
aruna@cdt.luth.se, lenka@sm.luth.se

Abstract

The layered video approach has been considered as one of the many approaches for solving the problems related to heterogeneity in the current Internet for real time applications. A number of schemes have been propounded taking into consideration only the network requirements. Another aspect is to develop layering scheme based on a combination of network requirements and the perceptual quality of video. In this paper a post-codec controller has been described that sends the video sequence in an optimal way taking into considerations the effect of type of video sequence, number of layers and distribution of frames among different layers on the perceptual quality of video, load on sender and bandwidth requirement. The controller has been tested in real network for two different configurations. (a) one sender and one receiver; and (b) one sender and many receivers connected in parallel. It has been observed that the quality of transmission improves by the use of the controller.

1. Introduction

As a result of the development of real-time multimedia applications the current network is under stress. Hence, not only is the performance of these applications often poor, it also affects the performance of other applications. This is because:

- The real time applications in general are less tolerant to the packet delays and losses.
- They require higher bandwidth for sending a sequence of high quality images.
- The uniform transmission of real-time media results in congestion of low bandwidth links and under-utilization of high capacity links.
- The quality of media is normally sensitive to dropped and delayed packets.

In order to solve the problems associated with the reliable multicast of video within a heterogeneous environment a number of approaches have been suggested. One of the promising methods is the transmission of video in layers [1-4]. A number of protocols have been proposed in order to send video in a number of layers. These protocols can be roughly divided into two main categories: sender-initiated and receiver-initiated. In the sender-initiated approach the sender multicasts a single video stream whose quality is adjusted based on the feedback information from the receivers. In the receiver-initiated approach the video signal is sent in a number of layers, where each stream uses a separate IP-multicast group address. The base layer contains the basic information of the image, and the enhancement layers provide further refinement to the basic quality of the image. At the receiver end, the receiver can subscribe to one or more layers based on its capability. The combination of these layers forms the complete video. Since, both the schemes have their respective advantages and disadvantages there are certain hybrid approaches that combine the good points of the both.

The proper way of division of video into layers is an important factor on which the success of the layered transmission would depend. For proper distribution it is important to decide:

1. The number of layers the video need to be divided into.
2. The distribution of available bandwidth among the layers.
3. The way – temporal, spatial or data partitioning by which the layering has been done.

To judge the suitability of scheme the test parameters should be:

1. The perceptual quality of video at the receiver end.
2. Computational complexity of the codec that determines the load on the sender and receivers.
4. The ability to satisfy the required heterogeneity.

A few protocols have been developed to take into account these factors. For example, Packet Pair Receiver Driven layered multicast protocol (PLM), which uses a packet pair to infer the available bandwidth at the bottleneck receivers to decide the optimum layer [6]. Source Adaptive Multi-Layered Multicast Algorithm (SAMM) [7] uses feedback from the receiver to the sender and a feedback merger sends the feedback regarding the network situation [8,9,10,11].
The knowledge of proper layering technique is of importance to the developers of Internet based conferencing tools, which are increasingly being used for distance education and health care [3]. In these applications, users located in geographically distant places are connected by links having varied bandwidth capacities. In the present network the private subscribers often have rather low availability of bandwidth. Hence, these tools work quite satisfactorily using rather "old" codec like Intra-H.261, which is designed for low bit rate transmission.

In order to solve the problems related to temporal synchronization, Intra-H.261 uses temporal compression model called conditional replenishment. Frames are made of blocks of 8 x 8 or 16 x 16 pixels. Only those blocks that change substantially in successive frames are encoded and sent. If a particular block undergoes only minor change it is not replenished and this causes “blockiness” in the picture. In order to take care of this problem “block aging” algorithm is used [4].

While dividing a video sequence into layers a number of factors need to be taken into account. These include:

- **Complexity of Video Sequence.** The architecture of the Intra H.261 is such that the codec sends only the changed blocks. Hence, as the changes taking place in a sequence increases the number of macroblocks sent per frame also increases. This puts additional burden on the sender and on the network and also affects the perceptual quality.

- **Number of Layers.** The increase in the number of layers increases the scalability, improves the video quality due to optimum utilization of bandwidth but it also increases complexity of the codec, and decreases codec efficiency. Therefore, increasing the number of layers follows the law of diminishing returns. This optimum number depends upon the type of codec and the partitioning method.

- **Distribution of Frames among Layers.** The distribution of frames among different layers affects the perceptual quality and the load on the network.

- **Frame Rate.** As the frame rate increases the perceptual quality of video also increases, but this happens at the expense of increased bandwidth requirement and load on sender.

The quality of video can be judged either subjectively or objectively. Till now there is no standard method for testing the quality of video transmitted over Internet. For subjective test, in the absence of any other standard, the ITU-R and ITU-T [7, 10] standards for digital TV may be used. While the subjective test may give the "ultimate truth", it has a number of disadvantages: it is time consuming and it is difficult because it is influenced by a number of factors like frame rate, lighting, image size, movement, degree of synchronization. There are a number of objective tests. Some of these, like peak signal to noise ratio (PSNR) and root mean square error (RMSE), are simple but do not correlate well with the subjective evaluation. Other more complex matrices are based on the complex spatial and temporal model of human visual system. These models evaluate the video much more accurately but are more difficult to work with.

The main idea of our study is to develop best layering technique for an inexpensive video conferencing tool suitable for general use in distance education and healthcare. For this purpose the effect of various layering parameters on the perceptual quality of video, bandwidth requirement and load on sender has been experimentally studied. From the results obtained an architecture has been developed that dynamically calculates the optimal layering technique. This paper describes the experimental results and the architecture of the dynamic layering controller.

2. Architecture of the System

The architecture has been developed for the optimal post-codec temporal layering in Intra H.261 codec, but with modifications it can be made more general and used for other codecs also. In this architecture a controller dynamically characterizes the nature of video sequence and takes input from the receivers regarding the network condition and then divides the video sequence into layers in an optimal way before transmitting it into network.

Figure 1 shows the block diagram for the architecture. The sender sends the video stream to an Intra-H.261 CODEC. The encoded stream goes to a demultiplexer, where the header of the RTP [14, 15] packets is read. On the basis of frame numbers the demultiplexer partitions the stream and sends them to different multicast addresses. At the receiver end, the multiplexer combines all the accepted layers and sends them to the decoder.

Figure 2 shows schematically the working of complexity calculator. In the complexity calculator the number of Macroblocks per frame is read for a set of frames. At the first instance the number of frames in a set is kept at six. Using the concept of linear prediction [18] the expected number of Macroblocks per frame for the seventh is calculated. This predicted value is used as the average complexity for the next set of frames. While the predicted value is being used by the controller, the next set of actual complexity is stored and used to calculate the next predicted value. The next set of frames has the same number of frames as the number of layers in which the video sequence shall be divided into.
Figure 1. Block diagram of the proposed architecture. Numbers 1-5 represent the complexity switch. Alphabets A-C represent various algorithms used for distributing the frames among various layers. Items in the box need to be optimized before the controller adjusts the demultiplexer.

Figure 2. Flow chart of the working of the complexity calculator when Algorithm A is used. F_x, F_{(x+N)} = No. of macroblocks in frame numbers x, (x+N) respectively. N = No. of layers in the previous set.

Using the scheme shown in Figure 3, the predicted complexity is used by the controller to calculate the number of layers and distribution of frames. The controller takes the predicted complexities and bandwidth-availability to the receivers as input. First using the co-relations between complexity - number of layers sent - number of layers received - bandwidth requirement, the available options of number of layers sent (1 to 6) - number of layers received (1 to 6) is found.

Next, using the co-relations between the complexity - number of layers sent - number of layers received - perceptual quality of video the best option based on the perceptual quality is selected. Similarly the best options for different receivers, having different bandwidth capacities, are calculated.
The relationship between the complexity of video, number of layers and distribution schemes is found empirically using experimental tests. The effect of these parameters on the bandwidth requirement is found both deductively and experimentally.

Using statistical analysis the scheme that will best satisfy the group of receivers is accepted. Depending upon the distribution profile of the receivers, different techniques may be adopted.

In order to have coordination between the sender and receiver, before changing the number of layers the sender first sends a message to the receivers regarding the change. The receiver then adjusts its existing position to a nearest bandwidth equivalent position to the new system by using the conversion database present with the receiver.

3. Experimental Procedure

Development of the Controller [19]. A filter was designed at the post codec level (a demultiplexer after the codec). Video from a VCR/DVD player was sent to the codec and subsequently to the demultiplexer. The demultiplexer read the timestamp of the frame, and then based on the algorithm the frames were divided into layers. In the present investigation the video were sent in 1 to 6 layers. Each layer of the coded video was sent on a unique IP multicast address. At the other end the receiver joined one or more addresses to receive the video [3].

Since the type of video sequence had significant effect on the compression scheme of the codec, hence, eight different types of video schemes were transmitted and tested. Their complexity was judged by counting the average number of macroblocks per frame. The simplest one was image of a stationary object transmitted continuously, it had a complexity of one macroblock per frame. At the other extreme was an image in which the scene was continuously zoomed in and out, having a complexity of 14 macroblocks per frame. In between these two limits were different kinds of video sequences, like people talking, sports, soldiers marching, conferencing.

Two types of algorithm were tested (Figure 4):

**Algorithm A.** It sent the successive frames equally on the different layers.

**Algorithm B.** In this algorithm the frames were sent on different layers so that more frames were sent to higher numbered layers.

\[ N_L \approx 2^{L-1} \]

In this scheme the base layer requires very less bandwidth but is able to take only the most essential data.

![Figure 4: Distribution of frames on different layers for Algorithms A and B.](image)

![Figure 5: The configurations used for testing the performance of the controller.](image)

Testing of the Controller. Two different types of configurations were tested (Figure 5). The sender was connected to a router by a link with a capacity of 100 Mbps. The router was a LINUX system in which one or more class-based differentiated services are implemented where the class based queuing was performed (CBQ). A video sequence of 2.5 minutes was sent though the set-up and the film recorded at the receiver end. The movie sequence was so chosen as to have sections having low (2 macroblocks per frame) to high (15 macroblocks per frame) complexity. Three different tests were carried out: (a) using Algorithm A, (b) using Algorithm B and (c) using both Algorithms A and B.

In the Configuration 1 there was one sender and one receiver. The receiver was connected to the router with a link having variable capacity (20, 40, 60 and 80 kbps). In the Configuration 2 there was only one sender but four receivers. All the receivers were connected in parallel and had link capacities of 20, 40, 60 and 80 kbps. The same video sequence as the previous case was used and the same three tests were carried out.
Testing of Perceptual Quality. The perceptual quality was initially tested using Peak Signal to Noise Ratio (PSNR), but was not found suitable. Thereafter, the perceptual quality of video was subjectively tested according to the recommendations of ITU-R standard. This standard was considered to be appropriate in the absence of any other standard made specifically for the conferencing tools. For each type of video sequence 15 people were asked to give their opinion. The subjects were all postgraduate students from various areas of engineering. They all had over 5 years of experience in computer and Internet operation. Audio was muted in the experiment.

During the development of the controller a slight modification was made to ITU-R [16, 17] standard so as to have quality scale between 1-10, instead of 1-5. This was done to make the judgment finer. The subjects were asked to consider the video grabbed just before the being sent to codec as having quality 10, and all other video sequence were judged in comparison to this sequence.

During the testing of the controller, as a reference the same movie was sent without any layering. This movie was given a value 0, and the test video sequence were judged in relation to this. The scale varied between +5 (better performance) to −5 (worse performance).

Statistical Analysis of the Experimental Data. Statistical analysis of the experimental data was carried out using a commercially available software. All the sets of experimental values had standard deviation less than 1.5. In order to judge that the difference between the reference video sequence and the test video sequence is significant the Wilcoxon Signed Ranks Test was carried out. In order to accept significant difference Z value should be less than −2 and P value should be less than 0.05.

Testing of Bandwidth Requirement. The bandwidth requirement was found by taking mean of the values of bandwidth (kbps) used at the receiver end of the conferencing tool.

4. Results and Discussion

Development of the Controller. Based on the experimental results a number of co-relations were developed between (a) complexity of video sequence, perceptual quality of video, number of layers sent and number of layers received; and (b) complexity of video sequence, bandwidth requirement, number of layers sent and number of layers received.

Algorithm A, \[\text{PQ} = 7.8 \left( \frac{\text{LR} + 3.1}{\text{LS} + 2.9} \right)\]

Algorithm B, \[\text{PQ} = 8 \times 2^{(\text{LR} - \text{LS})}\]

Algorithm A, \[\text{BR} = 80 \left( \frac{\text{LR}}{\text{LS}} \right)\]

Algorithm B, \[\text{BR} = 80 \times 2^{(\text{LR} - \text{LS})}\]
Figure 7. Improvement in the perceptual quality of video in the scale −5 to +5 on the use of controller. Unlayered video has a reference value 0.

Figure 6 gives an example of the results obtained. It shows the effect of number of layers sent and number of layers received for a video sequence having complexity 4 on the perceptual quality and the bandwidth requirement. Similar work was done for the video sequences having complexity 2 to 10. The equations so obtained were used in the development of the controller.

Testing of the Controller. It was observed that when the video sequence was sent without layering there were random packet drops depending upon the available bandwidth resulting in the deterioration of the video quality. Expectedly the lesser the bandwidth available poorer was the quality.

Only Algorithm A. For the Configuration 1, when the controller was run using only Algorithm A the subjects reported better performance for the video sequence with layering compared to video sequence without layering in the case of 40 and 60 kbps links (Figure 7). The subjects did not report any improvement in the performance for the video sequence with layering compared to the video sequence without in the case of 20 and 80 kbps. The lack of improvement in video quality upon layering in the case of 20 kbps bandwidth link may be because of two reasons. Firstly, the quality of the video sequence in both the cases was so bad that the subjects could not see any difference. Secondly, at higher complexity the bandwidth requirement even for one layer may be more than the available 20 kbps resulting in further drop of packets. On the other hand in the case of 80 kbps the quality of both was similar because there were no packet drops.

Both Algorithm A and B. When the controller used both the Algorithms A and B, the controller selected them automatically. Figure 8 is a representative example of the video sequence sent through the 20 kbps link in Configuration 1 and in Configuration 2. The figure clearly shows the switching between the two algorithms in the controller based on the complexity at a given instance. Comparison between the two sections of the figure shows that at lower complexity, hence, lower number of macroblocks per
frame, the controller sends the video in one layer using Algorithm A. At higher complexity, hence, higher number of macroblocks per frame, the controller sends the video in two layers using Algorithm B.

As Figure 7 shows, in this case the subjects reported better performance for the video sequence with layering compared to without layering in the case of 20, 40 and 60 kbps links. They did not report any improvement in the performance for 80 kbps link. This may be because the Algorithm B works better at the lower bandwidths. At the 80 kbps link again there were no packet drops, hence, no effect of layering.

5. Conclusions

From the work carried out it may be concluded that it is possible to make a simple post codec controller that can divide a video sequence into layers before sending to the network. By properly controlling the layering procedure it is possible to improve the quality of video at the receiver end by replacing random packet drops with more controlled packet rejections.

6. References


15. RFC 3158 “RTP Testing Strategies”, August 2001


Encoding of Video in Layers Based on Perceptual Quality of Video

ARUNA THAKUR AND LENKA MOTYCKOVA-CARR
Department of Computer Science and Electrical Engineering
Luleå University of Technology
SE – 971 87, Luleå
Sweden
aruna@cdt.luth.se, lenka@sm.luth.se

Abstract: - Layering techniques for video transmission are a promising way to satisfy receivers in heterogeneous environments. A number of receiver-initiated, sender-initiated, and hybrid protocols have been developed to transmit video in layers. These are supported in different ways by many CODECs. However, there is still a need to do a comparative study of various layering techniques to determine the best way. The study should include the effect of number of layers, bit rate and frame rate of each layer, available bandwidth and rate of packet drop on the perceptual quality of video. This paper examines some of these factors and describes our preliminary results.

Key-Words: - Video, layering, CODEC, perceptual quality,

1 Introduction
With the development of multimedia applications, like video conferencing tools, distance education and media-on-demand, the current network is under stress that not only gives poor quality for these applications but also effects the performance of other applications, This is because:
• The real time applications in general are less tolerant to the packet delays and losses.
• They require higher bandwidth for sending a sequence of high quality images.
• The real-time media are transmitted at a uniform rate to all the receivers in the network. Hence, the source must run at the bottleneck rate resulting in high-capacity links getting under-utilised; otherwise, it will overload portions of the network having low bandwidth capacity.
• The quality of media is normally sensitive to dropped packets. Retransmission of the lost packets is not a good solution, since, it may lead to NACK implosions and large-scale congestion. Also, retransmitted packets may arrive at the source after the time when they are required.

These problems are of special concern for the development of inexpensive desktop video conferencing tools, which are increasingly being used for distance education and health care. In these applications, users located in geographically distant places are connected by links having varied bandwidth capacities. For these applications to become successful they should:
• Be able to satisfy requirements of different types of users, a task that has become rather formidable with the proliferation of broadband on one hand and the wide spread use of mobile phones and Personal Digital Assistants (PDAs) on the other.
• Be able to provide reliable, good quality video because the users tend to compare the quality of the video with that of TV broadcast.
• Be affordable.

Since these conditions are difficult to meet the acceptance of Internet-based conferencing tools has been rather restricted. This is due to, first, poor technology; and, second, the inability of the users to use the tools to their maximum potential. While the second problem can be improved with training, the first needs to be solved at a more fundamental level.

In order to solve the problems associated with the reliable multicast of video within a heterogeneous environment a number of approaches have been suggested. These proposed approaches fall roughly into three categories:
1. Using a network capable of resource reservation to provide performance guarantees. Examples of this approach are resource ReSerVation Protocol (RSVP) and ATM [1]. These have not yet been deployed on large scale because:
   a. It is expensive to reserve bandwidth and hence only part of the video can be sent by this way. The rest needs to be sent by best effort.
   b. It is difficult to accurately estimate the bandwidth requirement for a particular application.
2. Using repair schemes employing interpolation or redundant data.
3. Using adaptive control techniques that can dynamically adjust the multimedia traffic characteristics to meet the network’s current capabilities.
A good solution is to efficiently partition the video into base and enhancement layers. The most important data is sent in the base layers that can be accepted by all receivers, and the rest is in the enhancement layers. The layering approach combined with priority packet dropping should be able to satisfy the restricted users and provide good quality to the others. For this scheme to be really successful and be accepted by the end-users, it is important that the division and transmission of video is not only based on the requirements of the network but also takes into consideration the psychological and physical performance of the human visual system.

Since achieving total satisfaction for all users is next to impossible, there is a need to establish not only the optimal quality but also the minimum and the maximum media quality required by a particular group for a particular task. The minimum limit would be the one which would be acceptable to the users, and the upper limit would be the point of diminishing return wherein the increase in quality does not necessarily mean increase in user satisfaction [2].

2 An Overview of Issues in Layered Video Transmission

2.1 Layered Transmission of Video

A number of protocols have been proposed in order to send video in a number of layers. These protocols can be roughly divided into two main categories: sender-initiated and receiver-initiated. In the sender-initiated approach the sender multicasts a single video stream whose quality is adjusted based on the feedback information from the receivers. In the receiver-initiated approach the video signal is sent in a number of layers, where each stream uses a separate IP-multicast group address. The base layer contains the basic information of the image, and the enhancement layers provide further refinement to the basic quality of the image. At the receiver end, the receiver can subscribe to one or more layers based on its capability. The combination of these layers forms the complete video. Since, both the schemes have their respective advantages and disadvantages there are certain hybrid approaches that combine the good points of the both.

The classical receiver-initiated solution that has been proposed is the Receiver-driven Layered Multicast (RLM) approach [3]. RLM operates within the Internet protocol architecture and multicasts as best effort, multi-point packet delivery. In this approach the sender transmits the video at constant rate, and the receivers dynamically judge number of layers they can receive by carrying out “join experiments”. While RLM has been hailed as a milestone in improving the scalability of video transmission, it suffers from certain disadvantages:

- Longer encoding and decoding time,
- Larger buffers at the receiver,
- Link congestion and inefficient convergence during join experiments,
- Misunderstandings about information shared among receivers,
- Unfair division of bandwidth among receivers.

Some of the problems faced by RLM have been solved in the Layered Video Multicast with Retransmission (LVMR) protocol [4]. In this approach, video quality has been improved by intelligently retransmitting lost packets and by the receivers dynamically adjusting the video reception rate using a hierarchical approach.

Multi-layered encoding alone cannot solve all the problems associated with the varying bandwidth constraints in multicast applications. Hybrid protocols like Source-Adaptive Multi-layered Multicast (SAMM) and Layered Multicast Control Protocol (LMCP) combine the strengths of sender-initiated and receiver initiated approaches [5,6]. Figure 1 is the generalised architecture for a hybrid layered video multicast. In these protocols, the sender sends the video in layers. The receiver accepts the layers depending upon its capacity and also sends a feedback to sender. Based on this feedback, the sender adjusts the number of layers and the transmission rate for each channel. The main advantages of these approaches are:

- The sender sends only the controlled number of layers, and does not unnecessarily overload the network.
- Transmission is more finely tuned to the requirements of the network.

All these approaches work on the philosophy of “best effort”, wherein the delivery of packets is not guaranteed. Hence, due to congestion some packets may get dropped. This drop in packets is normally uniform which means that packets may get dropped from any layer. Since the base layer contains the more sensitive information, it is not a good idea to let a packet from the base layer drop to adjust to the available bandwidth. A better solution can be achieved by using a priority-based approach wherein the priority is given to a layer depending upon its utility. This priority dropping may be achieved by the Differentiated Services (DiffServ) architecture.
These and many similar approaches may be good schemes for transmitting video in layers. But, the question that still needs to be answered is how the video should be partitioned. For this two aspects need to be considered:

- The number of layers the video need to be divided into. With increase in the number of layers it is possible to satisfy greater heterogeneity in the network. On the other hand the increase in the number of layers would increase the network overheads and complexity of CODEC (COder-DECoder). Hence, it is imperative that the number of layers be kept at a minimum.

- The distribution of available bandwidth among the layers. The base layer normally contains the most useful information. Ideally the major share of the bandwidth should be allocated to it. On the other hand, if a large part of the bandwidth is given to one layer, the user would not have a strong reason to subscribe to enhancement layers.

The most suitable way for partitioning the system should be judged based on the satisfaction of the users and the requirements of the network. To judge user satisfaction, tests should be carried out on the perceived quality of video. The network “friendliness” of the layering technique can be judged based on:

- Computational complexity of the CODEC, which determines the load on the sender and receivers.
- Efficiency of the CODEC in terms of compression performance over a bandwidth spectrum.
- The ability to satisfy greater heterogeneity.

Since the division of video into layers is done at the CODEC level, understanding the CODEC is vital.

2.2 Perceptual Quality of Video

The perceptual quality of video can be judged either subjectively or objectively. Since it is the users who would ultimately decide the quality of video, the subjective tests are considered the “ultimate truth”. The most widely used methods for measuring the subjective quality of speech and video images have been standardised and recommended by the International Telecommunications Union (ITU). ITU-R and ITU-T recommendations address subjective assessment of image quality over television systems and multimedia applications respectively. Almost all the subjective tests are time consuming and complex. This is because assessing the subjective perception of quality is a difficult task, since it is influenced by a number of factors like, frame rate, lighting, image size, movement, degree of synchronisation, and even clarity of audio. (In one experiment, subjects noted improvement in video quality when actually only the audio quality had been improved!) [1].

Quality of media delivery can also be expressed in objective measurements (e.g., packet delay and loss). Normally, it is difficult to find any strong co-relation between the subjective and objective assessments. Popular simple distortion measures such as peak signal-to-noise ratio (PSNR) and root mean-square error (RMSE) do not correlate well with perceived quality. This is because these simple tests do not take into account the viewing conditions and the psychological and physical behaviour of human visual system. A general objective model should take into account the various aspects of visual information processing in the brain. These aspects include:

- colour perception;
- the response properties of neurons in the visual cortex;
- temporal and spatial mechanisms;
- spatial and temporal contrast sensitivity and pattern masking.

A number of matrices based on a complex spatial and temporal model of the human visual system (HVS) have also been developed. While they are more general and accurate, evaluation is more complex [7].

2.3 Division Techniques for Video into Layers for Different CODECs

The layering can be done inside the CODEC or as a post-processing filter at the system level. In the latter case, a simple technique is used in which a filter parses through the output stream and reads the header. Depending upon the type of layering required, the packets are directed to the appropriate multicast group. At the decoder, a multiplexer sequences video data from the different multiplexer groups and sends a stream to a decoder where it is decoded.
While some of the CODECs support layering others do not. Layered transmission is not supported by the H.261 CODEC, but it can be done by modifying the CODEC or by adding a filter after CODEC. While the original H.263 does not support layering, the extended version of H.263 (H.263+ or H.263 Version 2) can support up to 15 layers. MPEG-2 supports layered representation but does not work efficiently at low bit rates because it relies on intra-frame updates to resynchronize the decoder in the presence of errors or packet loss.

Layering of video data can be done by means of any of the following four layering techniques. The performance of these layering techniques depends upon the CODEC used.

2.3.1 Temporal Layering
In this technique, it is possible to separate layers into Intra (I), Predictive (P) and Bi-directional (B) video frames. I-frames can be independently decoded, while P-frames require I-frames, and B-frames generally require both I- and B-frames for decoding. During congestion, the order of preference for the drop is first B-, then P- and finally the I-frames.

This type of layering is easiest to implement and introduces almost no overhead because it can be done at the post-CODEC level by reading the header to identify the type of frame. The problem with its performance is poorest since there may be conditions where dropping B-frames may result in poorer performance than the dropping P-frames [8-10].

2.3.2 Layering based on Data Partitioning (DP)
In this technique, layering is performed by allocating data in the bit stream (motion vector information and Discrete Cosine Transform or DCT coefficients) to different layers. Those vectors and coefficients with higher importance can be directed to the base layer and the rest to the enhancement layers.

After Temporal layering, partitioning based on DP is easiest to implement because it requires only a post-processor filter. The efficiency of this method of layering is also high since only headers (namely, sequence, group of pictures, picture and slice headers) are included in all layers. Unfortunately, it provides lower performance than the Signal-to-Noise Ratio (SNR) and spatial layering. It has been seen for MPEG-2 the quality of picture increases through 4 layers. Beyond this, there is not much increase. It has also been seen that even a small number of packet drops in base layer degrades the quality significantly while the quality falls almost linearly with the number of packets lost in the enhancement layers. Therefore, the base layer should be protected at the expense of the enhancement layers [9].

2.3.3 Signal to Noise Ratio (SNR) Layering
SNR layering is done at the CODEC level by encoding the video using a quantizer scale $Q_b$ in order to generate a base layer, and then generating the enhancement layer by encoding the difference between the original video and the base layer using a quantizer scale $Q_e$ [9]. For more than two layers, the same process is applied recursively using additional quantizer scale parameters.

In this type of layering the efficiency is rather low because the DCT coefficients are divided among the layers, and information for every DCT coefficient is included in all the layers. The efficiency falls rapidly with an increase in the number of layers, so it is not practical to have more than 2 layers. For SNR layering, the quality degrades linearly with the rate of packet loss in enhancement layer and drops rapidly when the data loss occurs in the base layer [9].

2.3.4 Spatial Layering
This scalability provides the best performance, especially when network conditions are poor, but has the highest implementation complexity. Perceptual quality increases almost linearly with the increase in number of layers, and efficiency decreases almost linearly with the number of layers [10].

3 Our Work
In the on-going work, we are determining how the video streams should be divided into layers to get the best perceptual quality. We are examining different network conditions in order to judge the suitability of video quality delivered by individual CODECs and layering procedures.

3.2 Experimental Technique
We carried out the study using a commercially available video conferencing tool – Marratech Pro. Marratech Pro has a number of features like real-time video, audio, shared application windows, whiteboard, chat and web-based presentations. To send and receive video it uses the Intra-H.261 CODEC, and the video is then packetized and sent through RTP streams. The user can directly adjust the frame rate (frames per second), the quality or “blockiness” of the image, and the overall bandwidth used [11].
First, we evaluated the video quality of the conferencing tool. The effect of frame rate and image quality on the perceived video quality was subjectively judged. Fifteen university employees, from three different departments, and having varied education level (2 years college to Ph.D.) participated in the experiment. Subjects’ experience in use of video conferencing ranged from no experience, to more than 5-year’s. Subjects were asked to rate six video recordings (50 seconds each). Audio was muted in the experiment.

In the experiments the independent variables were frame rate (8 or 18 fps) and image quality (8 or 18 in arbitrary units). Bandwidth (128 kbps) and packet loss (1%) were kept constant throughout the experiments. Limited number of experiments were also carried out using high bandwidth (700 & 1024 kbps), frame rate (21 & 25 fps) and image quality (26 & 30).

When the bandwidth was not a constraint, as expected, subjects preferred to have high frame rates and high image resolution. For limited bandwidth, users preferred high frame rates to high image quality.

Next, we modified the conferencing tool at the system level to include a filter (a demultiplexer written in Java) after the CODEC. Figure 2 shows the architecture for video transmission. The sender sends the video stream to an Intra-H.261 CODEC. The encoded stream goes to a demultiplexer, where the header of the RTP packets is read. On the basis of macroblock address the demultiplexer partitions the stream and sends them to different multicast addresses. At the receiver end, the multiplexer combines all the accepted layers and sends them to the decoder.

Figure 3 shows the effect of the layering technique on the quality of video. When only the base layer was received the quality was very poor compared to the quality of video when both all layers were received.

### 3.3 Discussion

Based on the preliminary investigation a generalised diagram (Figure 4) has been generated. It shows the effect of number of layers on various parameters.

- **Coding efficiency (c):** Can be defined as the ratio of the number of bits used by the non-layered coder to the number of bits used by the layered CODEC to achieve the same quality of picture. In some cases, the efficiency may fall drastically with the number of layers \[9\].

- **Coding complexity:** Is the additional calculations that must be performed by the CODEC in order to partition the video into layers. Increasing complexity the coder and decoder would require additional time, resulting in poor video quality \[9\].

- **Ability to handle heterogeneity:** Is the upper and lower limit of bandwidth that can be satisfied by the layered stream. By increase in the number of layers, it is possible to satisfy greater heterogeneity, but the range of bandwidth covered does not increase linearly with the number of layers \[3\].

- **Perceptual quality of video:** Normally increases with more layers, but the quality does not increase linearly. It is even possible that the quality may fall with an increase in number of layers because of increased complexity and inefficiency. This is especially true for layering based on SNR due to its bad scalability \[8,9\].

The increase in the number of layers not only improves the quality but also increases complexity of the CODEC, and decreases CODEC efficiency. Therefore, increasing the number of layers follows the law of diminishing returns. Beyond a certain number, any further increase in the number of layers would make the overall quality worse. This optimum number depends upon the type of CODEC and the partitioning method.
3.4 Future Work

The main idea of the project is to develop best layering technique for inexpensive video conferencing tool suitable for general use in distance education and healthcare. In future, the effect of layering techniques, the number of layers and the frame rate and bit rate of each layer on the perceptual quality of video will be studied. From this investigation, a function will be developed to calculate the perceptual quality of video given the layering technique, number of layers, available bandwidth and its distribution among different layers, frame rate and bit rate of each layer.

It is also assumed that for a video conferencing tool the head and the shoulders of a person are more important than the background. Therefore, the area of the frame containing the head and the shoulder should be given higher priority and be sent to the base layer. The rest of the image would be sent to an enhancement layer. Efforts are being made to divide the video into layers in a context dependent way using conditional replenishment where only the blocks that have changed are encoded and sent in succeeding frames.

References: