

TelSign:

Telecommunications and Sign Transmission

Final Report to DTI March 1997
An investigative project
on the possibilities for
sign video transmission in mobile telecommunication

Prepared by the
Centre for Deaf Studies,
University of Bristol

Jim Kyle, Pete Hinchliffe,
Mick Canavan, Matt Page, David Jackson

A high priority for deaf people in telecommunications is the capability to send and receive signed messages. Solutions to the problem of meeting this bandwidth demand have been patchy and incomplete even with ISDN 128Kbps lines. Experiments with Picture-Tel systems in CDS indicate that sign language is readable in certain conditions. In this project, the prospects for sign transmission in mobile telecommunications have been explored by examining the literature and published project work and by conducting studies of sign transmission and reception using currently available systems. Results indicate problems for some signing conditions even though the technical outlook is good.

Table of Contents

CHAPTER 1: DEAF PEOPLE AND SIGNED COMMUNICATION.....	5
<i>Deaf People</i>	5
<i>Mobile Communications</i>	5
CHAPTER 2: VIDEO ON THE INTERNET.....	7
<i>Introduction</i>	7
<i>Internet Protocols</i>	7
The IP.....	8
TCP.....	8
UDP.....	9
Other Protocols.....	9
<i>Summary</i>	10
<i>Commercial Applications for Encoding and Viewing - 'Streamed Video'</i>	10
Alaris Products: http://www.alaris.com/	10
Iterated Systems: http://www.clearvideo.com/	11
Progressive networks: http://www.realaudio.com	11
VDOnet: http://www.vdo.net	12
Vxtreme: http://www.vxtreme.com	12
VivoActive: http://www.vivo.com	12
Vosaic Corp: http://www.vosaic.com	13
Xing Technology Corporation: http://www.xingtech.com	13
<i>Conclusions</i>	14
CHAPTER 3:.....	15
DESKTOP VIDEO-CONFERENCING.....	15
<i>Latest developments in H.324 Compliant Solutions for Video Conferencing</i>	16
<i>Conclusion</i>	Error! Bookmark not defined.
CHAPTER 4. VIDEO COMPRESSION.....	18
Compression Requirements for Sign Language.....	18
<i>Current Compression Techniques</i>	18
Videotelephony.....	19
<i>Compression for Internet Applications</i>	19
Bandwidth Scalability.....	19
Resolution, frame-rate, frame-quality scalability.....	19
Fast compression/decompression.....	19
Ability to cope with network losses.....	20
Encoding and decoding latency.....	20
<i>Second Generation Video Compression Techniques</i>	20
Model Based Coding.....	21
Frame Segmentation.....	22
Advanced encoders using new techniques.....	22

Promising results.....	22
CHAPTER 5: POSSIBILITIES AND SPECIFICATIONS FOR VIDEO.....	23
<i>MPEG4</i>	23
<i>H.263L</i>	25
<i>Submissions for H.263L</i>	25
CHAPTER 6 - TRIALS.....	26
<i>Performance of H.263 at low bit-rates:</i>	26
<i>Preliminary Findings</i>	27
<i>Informal Tests</i>	27
<i>Sign Language and Video Conferencing Trials</i>	28
<i>Test 1</i>	28
<i>Conclusions</i>	29
<i>Test 2</i>	30
<i>Conclusions</i>	31
<i>Test 3</i>	31
<i>Test 4</i>	32
<i>Discussions</i>	32
CHAPTER 7 - THE WAY FORWARD AND SPECIFICATIONS.....	33
<i>Requirements and Advantages in Sign Communication</i>	33
<i>Current Capabilities</i>	33
<i>Implications</i>	34
<i>Future Needs</i>	35
CHAPTER 8: SUMMARY & CONCLUSIONS.....	36
<i>Currently Available Video Technology</i>	36
<i>The Future</i>	36
APPENDIX 1.....	37
<i>References</i>	37
<i>Compression Standards</i>	37
<i>Relevant ACTS Projects</i>	37
<i>Relevant documents</i>	38

Chapter 1: Deaf People and Signed Communication

Deaf People

It is perhaps ironic that in the 19th century, deaf people had better access to telecommunications than they do now. The first long distance messages were sent by telegraph and the position of deaf people was exactly the same as that of hearing people. Text messages could be sent and received. The further oddity is that the inventor of the telephone, Alexander Graham Bell married a deaf wife and through his invention prepared the way for over 100 years of discrimination against deaf people in the workplace - because they cannot use the voice telephone.

It was not until the mid-1970s that it became possible for deaf people to send text messages in live mode and to interact through text terminals. Even so, these text terminals are simplistic and expensive and are only feasible in the USA where local calls are not timed and are effectively free. Elsewhere deaf people have missed out on a communications revolution which has significantly enhanced the reach of hearing people and provided the means for social and economic interaction.

Although not a large group, deaf people are a significant number of people - nearing 500,000 in the European Union; and if those who lose their hearing later in life are included the numbers increase by tenfold. However, attempts to deal with their telecommunications needs have led to national solutions and incompatible standards. Deaf people may be able to type to each other in one country but differing standard make it almost impossible across national boundaries. There have been some advances in text communications provision with BT offering a relay service for text to voice and vice versa and also making special discounts available to deaf people. Text terminals have become more compact even though they are still analogue in nature with primitive single line displays. There are now some call stations in prominent public places where a deaf people can call by minicom. There has been a growing awareness of deaf people's needs.

However, when research has been carried out, deaf people express the desire for video communications rather than enhanced text communication. Deaf people would like to be able to use sign language directly with other deaf people. The difficulty is that the telephone system which is now in place does not have sufficient capacity for this possibility - at least not in its simple form. Software and hardware advances have tried to improve bandwidth and throughput of messages. Although pictures and to some extent video can be transmitted we are not yet sure that the picture clarity, resolution, and transmission rate is sufficient for deaf people's interaction at a distance. This project was set up to find out the answers to this question.

Mobile Communications

In the progress towards full mobile communications using radio telephony, the demands of pictures and video have not been ignored. A priority has been to develop a mobile terminal capable of handling internet communications as well as text messages. A considerable amount can be achieved with a 9.6kbps channel in terms of back-to-back text conversation and use of short message services (see TCALL project). Connected to a notebook computer, considerable flexibility can be obtained in terms of the type of data which can be transmitted and the software which can be used. The possibilities for video handling may be introduced as long as the stability of the radio transmission can be guaranteed at realistic cost levels. At the moment, this is not completely certain. However, in this project assumptions may have to be made about the future viability of the transmissions and that there will be available multi-slot transmission which will greatly enhance bandwidth. While

Ericsson have tested 64Kbps, it is proposed that 1Mbps services will be available in the not too distant future¹.

While video pictures and moving video can be sent through the Internet, there are still problems for high bandwidth applications such as signed language transmission. Although we believe there to be in-built redundancy in signing which may help transmission and reception, the requirements in a two-dimensional medium are not yet well established. Even television broadcasting of signing is problematic where information density is high. There remains a great deal to be done to determine the effective conditions for broadcast signing. Given the nature of the current bandwidth restrictions, the questions are then framed in terms of how much compression is possible of the video signal while still retaining the accurate perception of the signing. In our case, the questions are all linked to the needs of deaf people for person to person signing and the review must come to conclusions about how realistic this goal is within the latest developments in the field.

A reference point for this work is *DICTUM* which is an MPEG encoded sign language teaching CD which was developed at CDS in a EU TIDE project (McEntee, 1996). The CD has a curriculum for sign language involving a large number of video clips. These are viewed using a hardware or software based MPEG playback system and this gives us a video standard to which sign transmission in live conversation can be compared.

A set of simple questions have been set for this project and they can be summarised as questions about

- MPEG compression algorithms
- Transmission characteristics
- Requirements for sign production and reception in live conversation
- Video possibilities

Our starting point is the location which is most likely to deliver video information - the Internet. Because the field is changing very quickly publication of information is not able to keep up with the changes. New developments in video transmission are most likely to appear in research and product sites on the Internet. At the same time, it is the internet which has focused the effort on compression because the current bandwidth on its own, is hardly able to support video messages. This is set to change as video conferencing progresses and more people come to expect to see other people as well just to hear them on the phone. In order to make this a reality, considerable compression is required and this has been the focus of the MPEG4 standard development and the video conferencing specifications.

This project has examined the use of video in a number of telecommunications and information technology situations. Experimental work was carried out with deaf participants and this is also reported. Final conclusion are drawn and recommendations are made.

¹ CCR in Bristol are working with 1Mbps wireless LANs and HP are marketing 2Mbps wireless LANs. These may form test beds for future video applications.

Chapter 2: Video on the Internet

Introduction

Although there are material differences in the use of mobile telecommunications and in the access to the Internet, the obstacles to be overcome in sending video in the World Wide Web to modems with speeds of 28.8kbps, are not dissimilar to those which will be faced in transmitting and receiving video in the mobile arena. It is also true that once deaf people are given the facilities of a mobile terminal they will wish to utilise the vast resource of the internet and that this may be an on-line information source of considerable value. On these two counts, it makes sense to begin the review at this point.

To understand the difficulties faced by application developers in bringing real-time video to the Internet it is necessary to have a basic understanding of how the Internet works. The Internet is a 'packet switched' network which relies on the use of protocols. Protocols impose a set of rules which determine, among other things, how the data is to be packaged, how it is to be sent, and what sort of error correction is to be used.

This chapter will outline the current protocols which are in use today, beginning with the 'heart and soul' of the Internet, the TCP/IP (Transmission Control Protocol/Internet Protocol) protocols. It will then explain why these protocols are inadequate for the delivery of real-time video.

In an attempt to overcome the limitations of the TCP many applications use the User Datagram Protocol (UDP) instead. However, results have proved to be disappointing, at least as regards video (audio streaming has however made tremendous advances). This is not entirely due to the problems associated with the underlying protocols since 'network congestion' is a major factor.

It is widely accepted that the Internet is unable to deliver real-time video, of an acceptable quality, if current protocols continue to be used. There are now a number of emerging protocols that seek to combat the problems associated with the delivery of time critical data. In particular, many of these protocols specifically address the problems associated with 'network congestion'. The Internet Engineering Task Force (IETF) is a standards body which has published a number of RFCs (Request For Comments) which define these new protocols. Towards the end of this chapter we discuss the development of these protocols and explain just how they will improve upon current video transmissions. In particular, the protocols we will look at will be:

the Resource Reservation Protocol (RSVP),
the Real Time Protocol and Real Time Control Protocol (RTP/RTCP)
and finally, the Real Time Streaming Protocol (RTSP).

Internet Protocols

The difficulties in providing 'real-time' or 'streamed video' over the Internet are immense. There are numerous problems which relate to bandwidth limitations, network congestion, unsuitable compression techniques, unacceptable rates of 'packet loss' and the evolution of myriad different network technologies that are required to interact with each other.

The Internet was initially developed on a military agenda. The aim was to develop a communication network that was robust enough to withstand a nuclear war! Hence the growth of a network based upon the TCP/IP² (Transmission Control Protocol/Internet Protocol), which answered the needs of the military perfectly. The TCP³/IP suite of protocols are lossless transmission protocols which ensures that all data sent arrives at its destination. There is no guarantee of how long it will take or by which route the data will travel. TCP/IP strength lies in the fact that in its journey from A to B data can and does,

² Full specification of Internet Protocol Version 4 can be found in appendix 2.

³ Full specification of the Transmission Control protocol can be found in appendix 2.

take numerous paths to get to its destination, paths that cannot be mapped out in advance. These paths are by no means the shortest possible. A client requesting data from a server within his or her home town will probably be unaware that both the request and the delivery of the data has taken paths through a host of different countries.

It is already evident that the Internet was not designed for the delivery of multimedia type data. In the quest to deliver 'streamed' video over the Internet two related problems have already arisen. The data does not take the shortest path between A and B and the TCP/IP protocols are not concerned with issues relating to time; in fact, time critical data has no relevance to these protocols. TCP/IP provide support for relatively simple distributed applications, such as e-mail, file transfer, and remote access and it does this extremely well.

The IP

The IP is responsible for basic network connectivity, it is concerned with network addresses. The current standard is Ipv4 which is increasingly becoming overwhelmed with the changing needs and popularity of the Internet and a new IP is being developed. The fact that Ipv4 uses a fixed 32 bit address length means that in the very near future there will be a shortage of network addresses. Furthermore, and more relevant to this paper, the current Ipv4 has no support for real-time traffic, for congestion-control schemes or for security. In July 1992 the Internet Engineering Task Force (IETF) began looking for a new IP standard to replace the existing standard. In January 1995 it published RFC (Request for Comments) 1752 which outlined a new IP protocol officially known as Ipv6⁴. This version, Ipv6, uses 128 bit addresses and solves, at least for the foreseeable future, the problem of shortages of network addresses. However, IPv6 also simplifies and speeds up router processing by using a more efficient way of labelling packet headers. More importantly, IPv6 can label packets which belong to a particular traffic flow. This means that specialised traffic such as real-time video will receive special handling which should improve the quality of video. The introduction of Ipv6 will certainly help in the development of quality real-time video on the Internet but is unable to do this alone. We now need to look at the other half of the equation, TCP, and determine what relevance it has.

TCP

TCP is primarily concerned with error checking and sequence numbering. It establishes a dialogue between the computer sending the information and the computer that is receiving the information. This dialogue is said to be 'connection orientated' because if a packet is lost, TCP tells the server to re-send that data. It is a very efficient way of ensuring that *all* data requested is received. However, this protocol is not suited for the delivery of real-time or 'streamed' video, especially over the Internet which is notorious for packet loss. If a video file is requested by a client over the Internet and TCP is being used it is extremely unlikely, perhaps impossible, that the video will be smooth. It will almost certainly be 'jerky' and resemble a slide-show with the application continuously pausing as it waits for a lost packet to be re-transmitted.

A further problem with TCP and its suitability for video delivery is related to its initial design goals. TCP always tries to maximise data transfer and since it has no way of knowing what the available bandwidth is at any given moment, the data rate is constantly changing. TCP starts by sending data at a very low bit-rate and it increases this bit-rate until there is indication of packet loss from the receiving computer. TCP then reverts to sending very low bit-rates and begins the whole process again. What this means therefore is that a 28.8Kbps modem (probably the most common type of connection to the Internet) is repeatedly being driven to the point where it loses packets of data. In this scenario, smooth video reproduction over the Internet is not possible. However, the TCP/IP protocol is, as mentioned earlier, a suite of protocols and does not consist of just these two. Within this

⁴ Full specification of Internet Protocol version 6 can be found in appendix 2.

suite there exists a protocol called UDP⁵ (User Datagram Protocol) which sits on top of IP and can be used instead of TCP.

UDP

UDP provides no error checking or sequence numbering but is an efficient way of sending large amounts of data over the Internet. If a packet is lost UDP is not concerned and continues to send data as if nothing has happened. Now for video, assuming that the rate of packet loss is not too high, we can afford to lose some data packets and still receive video that is of fairly good quality. What is certain is that by using UDP the video will certainly be much smoother than video received via TCP. Furthermore, client applications themselves may provide for error checking and sequence numbering which are more suited to video transfers than TCP.

Other Protocols

Other protocols have also been developed, or are in the process of being developed, which will greatly improve the quality of video transmission and reception on the Internet.

The Resource Reservation Protocol⁶ (RSVP) is another protocol that sits on top of IP. RSVP allows the client application to request a particular Quality of Service (QOS) for its data. RSVP instructs routers along the data path to maintain a particular QOS and in effect manages to allow a router based network to work very much like a circuit switched network, i.e. on a best effort basis. Routers that do not support this protocol do not pose a problem as RSVP provides for transparent operation. RSVP works on any physical network as long as the underlying protocol is IP and RSVP is intended to work with Ipv6. The combination of these two protocols will allow users to set up 'end to end' connections with a specified amount of flow control for a given time.

Other protocols are currently being developed by the IETF which will aid real time data delivery on the Internet, two of these are the Real Time Protocol⁷ (RTP) and closely linked to it, the Real Time Control Protocol⁸ (RTCP). RTP can be used for media on demand or for services such as Internet telephony. RTP provides for timing reconstruction, loss detection, security and content identification and the aim is to make RTP very flexible so that it is not tied to any specific underlying protocol such as IP. RTP can be used with or without RTCP although without it there is no guarantee of Quality of Service (QOS). RTCP offers QOS feedback and support for real time conferencing on the Internet for large groups. RTP and RTCP are in an experimental stage although some applications have been developed and are RTP compliant. RTP may also be used alongside RSVP.

Progressive Networks Inc and Netscape Communications Corporation have submitted an RFC to the IETF concerning a protocol called the Real Time Streaming Protocol⁹ (RTSP). This is an application level protocol to control single or multiple streams of continuous time synchronised data.

Finally, Vosaic Corporation have developed the Video Datagram Protocol¹⁰ (VDP). This runs on top of the IP and is an application control protocol. It allows for a connection to be made between the client and the server, allowing the client to have some control over the video being received, i.e. pause, fast forward etc. It is an adaptive protocol in that it adjusts its' data rates in response to available bandwidth, thus is well suited to the internet environment. When a connection is made using VDP two channels are used, one is for unreliable data transmission and uncritical feedback and the other is the control information

⁵ Full specification of the User Datagram Protocol can be found in appendix 2

⁶ Full specification of the Resource Reservation Protocol can be found in appendix 2

⁷ Full specification of the Real Time Protocol can be found in appendix 2

⁸ Full specification of the Real Time Control Protocol can be found in appendix 2

⁹ Full specification of the Real-Time Streaming Protocol can be found in appendix 2

¹⁰ I was unable to find any specification for this protocol which may mean that it has yet to be submitted to the Internet Engineering Task Force.

which is sent from client to server. It is claimed that this protocol reduces inter-frame jitter of video reproduction and adapts to the clients CPU load.

Summary

New commercial applications are continuously being developed in response to the needs and wishes of the Internet community. Companies are attempting to provide both software and hardware solutions to the problem of delivering time critical data over the Internet. Some of these applications use TCP, UDP or proprietary protocols that run on top of IP to deliver video and audio. However, it is a fair assumption that time critical data will not flourish on the Internet until protocols such as RTP, RSVP and RTCP mature and become widely implemented. Once these protocols are in place companies can then begin to tackle effectively other problems associated with the delivery of time critical data.

Within a very short space of time the Internet has developed from a network intended to deliver text based material into a 'multimedia rich' environment. It has captured the imagination of the public and it is they who are continuously dictating the agenda. No longer satisfied with static images upon a World Wide Web page the public are awaiting commercial products that will deliver real time video. The next aspect to consider are these commercial responses to the problems.

Commercial Applications for Encoding and Viewing - 'Streamed Video'

This section reviews selected commercial applications currently available on the Internet. They will be reviewed in alphabetical order by company name. That is, we have made no specific judgement of effectiveness nor do we rank them in any sequence. The intention is merely to give information on the range of products which might be used now.

Alaris Products: <http://www.alaris.com/>

Alaris have produced an encoder entitled 'The Videogram Packager' to produce files with their own proprietary .VGM extension. It is a quick capture video card and full details regarding cost, specification and minimum computer requirements can be found in appendix 2.

Alaris claims that video files produced with its' proprietary codec are 2-4 times smaller than files produced with Indio, MPEG or CinePak codecs. Furthermore the codec is scaleable and can produce files with different data rates.

The Alaris 'Streaming Video Player' is available from the web site and is free.

Alaris uses basic TCP/IP to deliver video so that the user is sure of receiving the complete file. However, in light of the problems associated with TCP the player will continuously pause whilst the application waits for a lost packet to be re-transmitted. However, the company is not selling this product as a real-time streaming video player and is instead concentrating upon its possible use for sending video grams via e-mail. Once a video file has been 'packaged' using the 'Videogram Packager' it can either have a file extension of .EXE or of .VGM. If it has the .EXE extension it means that the Videogram Player is included along with the video file. The player is only 100KB in size and allows for videograms to be sent to anyone who has a mail client that can accept attachments. VGM files are video files which do not include the player.

Files produced by the Videogram Packager were downloaded and tested¹¹. The first file was 1,651KB in size, resolution was 160x120, colour depth was 24 bpp and it achieved 15 frames per second. The Length of video clip was 82 Seconds.

¹¹ Tests were conducted on a computer with an AMD 100 MHZ processor. This is not a particularly 'fast' computer and probably represents a fair average of the processor power currently being used by the Internet community. Furthermore, all tests were conducted between the hours of 4.30am and 5.30am during weekdays when network congestion is at its least. I was using a US Robotics Sportster 28.8Kbps modem

At 15 fps the video was smooth although the quality was a little degraded. Facial features and details were not easily discernible although the quality of the audio was very good, synchronisation between audio and video was acceptable.

The second file was 449 KB in size, resolution was 160x120, colour depth 24 bpp and had 15 frames a second. The length of the clip was 20 seconds. This video had much more contrast than the previous one and detailed features were discernible. Again, the video was extremely smooth and synchronisation between audio and video was good.

Iterated Systems: <http://www.clearvideo.com/>

Iterated Systems has developed 'ClearVideo' which is a software based codec for compressing standard Windows (.AVI) and Quicktime (.MOV) movies. During the process of compression these files are changed into a format which allows for instant viewing, i.e. streamed video. To view these files *Iterated* have also developed a plug-in for Netscape Navigator which is free.

ClearVideo uses a compression technique developed from its fractal compression techniques used for still images. The company claims that using this compression technique results in fewer 'key frames' and therefore, smaller, more efficient video files. Full specification and minimum computer requirements for running their software based codec can be seen in Appendix 2.

Progressive networks: <http://www.realaudio.com>

Progressive networks approach the problem of delivering streamed video by creating the Real Audio protocol and developing a client-server architecture. The encoder and player are free but the server software is charged for. The price of this software depends upon the number of simultaneous streams that the server software will support.

Progressive Networks uses a new transfer protocol called the Real Time Streaming Protocol (RTSP) which it developed in partnership with Netscape Communications. This protocol supports bi-directional communication between clients and servers, enabling the client to pause, fast forward, rewind and skip to particular sections or tracks. RTSP sits on top of TCP/IP and both these protocols may be used; however, it is recommended that the UDP protocol be used in the vast majority of cases. Progressive Networks also claim to have developed a sophisticated loss correction system to cope with packet loss. This is done by using a system called 'Forward Error Coding', in essence this system 're-creates' the lost packet at the client's end and it is claimed that degradation of video quality is minimised. The necessity of this technique lies in the fact that generally streamed video over the Internet is possible because of, among other things, the low number of frames used per second. However, using a low frame rate means that each individual frame has to be rendered for a correspondingly longer period and the loss of data packets is therefore more damaging. Progressive Networks uses two codecs, 'Real Video Standard' and 'Real Video Fractal' the latter having been developed by Iterated systems as mentioned earlier. Both codecs are scaleable for all bit rates and it is recommended that the standard codec is used for networks where packet loss is expected to be high.

Using the dedicated server software means that data is delivered in a paced fashion, unlike an HTTP server which attempts to send data as fast as possible, hence there is less likelihood of packets being lost. Furthermore, server based software enables bandwidth negotiation, dynamic connection management and buffered play.

In order to test Progressive Networks streamed video files, we visited the web page entitled *RealVideo demos*. There are various options for downloading over a 28.8 Kbps link. First we chose a 'newscast' video quality clip, streamed at 20Kbps at 10 fps. The player is spawned on selection of the clip and has VCR functionality, i.e. fast forward, rewind and pause capabilities. There is a small delay of a few seconds due to 'buffering' which simply

and during the test period it continuously achieved data rate downloads which were close to its theoretical maximum.

means that the clip will not start playing until enough data is in the cache (the size of this cache can be determined by the client in preferences).

Whilst the streaming rate was fairly constant, due to the fact that the UDP protocol is being used instead of TCP, video quality was fair at best. It was fuzzy, a little blocky and resolution was poor. The stream came in at an average of 20.5 Kbps and although the expected frame rate was 10fps the average turned out to be approximately 8.7fps.

The second clip chosen was in 'buffered play mode' which meant that the whole file was downloaded before it began to play, presumably using TCP instead of UDP. Again, video quality was fairly poor with a lack of resolution but the video was fairly smooth in playback. For details¹² concerning P.C. requirements, cost and performance capabilities see appendix 2.

VDOnet: <http://www.vdo.net>

VDO also delivers streamed video through a client/server architecture and is in many ways akin to the approach used by Progressive Networks. In fact there is very little to choose between them, both use the UDP, both use scaleable compression algorithms (VDO using its' proprietary VDOWave codec) and both charge for their server software on a 'per stream' basis. The VDO product line consists of the VDOLive On-Demand Server, VDOLive Tools and the VDOLive player.

The VDO player is basic and the user has the options to play pause or stop. In tests, the video quality was found to be jerky, with poor resolution. It was not possible on either of the two clips which were downloaded, to discern the details of the face or hands. When downloading clips a small 'bar indicator' informs the user of the percentage rate of reception. Whilst this is not very informative, what is clear is that once this percentage drops below 80% the video stalls. Further information on their products, costs, P.C. requirements and VDO capabilities, can be found in Appendix 2

Vxtreme: <http://www.vxtreme.com>

Vxtreme has developed the Web Theatre platform which consists of the Web Theatre Producer, the Web Theatre Server and the Web Theatre Client. The approach used here is almost identical to the other two server/client solutions provided by Progressive Networks and VDO. The Web Theatre Producer is a software based video capture and compression tool which has a maximum compression ratio of 500:1 and claims that their codec is far superior to MPEG codecs. A novel approach which distinguishes Vxtreme from the other products reviewed in this report is the ability to synchronise the video with other web page elements. An example of this can be seen at Stanford Universities web site,

<http://tempest.stanford.edu/~sitn/cs244a/>

At this site, the video clip (including audio) is on the left hand side of the screen and consists of a lecture being given. On the right hand screen, notes continuously appear to support what is being said. When the lecturer writes upon the blackboard, an image of a blackboard appears on the right hand screen. Whilst Stanford university have used Vxtremes' technology in a very intuitive and effective way, once again the video quality is poor. The image is small with poor resolution although by using the UDP protocol Vxtreme do succeed in providing smooth video. Whilst the player offers similar functionality to those provided by Progressive Networks, i.e. the ability to fast forward etc., it is not as well developed or as stylish as its' competitors. For more details concerning cost, PC requirements and performance see Appendix 2.

VivoActive: <http://www.vivo.com>

VivoActive is a serverless based solution to streaming video on the www. It uses standard videoconferencing standards, namely H.263 for video compression and G.723 for audio

¹² Appendix 2 has been prepared on disk since it contains a large number of extensive entries. A table of contents is also provided on the disk.

compression, and uses TCP for delivery. The user has no control over the video stream except to start it, pause it or stop it. In short, this is a very cheap way of bringing streamed video to the www as it requires no server software. The quality of this streamed video is very poor, extremely low resolution, fuzzy images, jerky video and long pauses. For more details see Appendix 2.

Vosaic Corp: <http://www.vosaic.com>

Vosaic claim that the use of the Video Datagram Protocol (VDP) results in a forty-four-fold increase in received video frame rate compared to the use of the TCP protocol. A frame rate increase from 0.2fps to 9fps. VDP is a control protocol which allows the client to have some control over the data stream from the server. The VDP uses an adaptive algorithm which adapts to available bandwidth and to the clients CPU load. Vosaic video files can be created using either hardware or software from MPEG, AVI or Quicktime files. Vosaic supply a free plug-in for Netscape which operates in tandem with the Vosaic video server. The plug-in features real time download and display of:

- MPEG 1 and 2 video
- MPEG layer 1 and layer 2 audio
- H.263 video
- GSM and half rate GSM audio
- G723.1 and half rate G723.1 audio

The Vosaic plug-in proved to be disappointing. The clips within the 'video library' are separated into three categories; high bandwidth, medium bandwidth through an ISDN line and low bandwidth through POTS. The low bandwidth clips are all MPEG movies delivered with a maximum frame rate of 6 fps. Given this low frame rate one would expect good video quality, however, the resolution was extremely poor and each frame was extremely 'blocky'.

Vosaic Corporation are currently involved in experiments which involve the use of the Vosaic/ Mosaic browser. The Vosaic browser, is a radically different approach to providing real-time video/audio over the Internet. The Vosaic media browser allows for video to be displayed within a standard web page without the use of an external player, hence allowing for video menus. The Vosaic browser is configured to use TCP for text and image transmission whilst real-time video/audio playback uses VDP. Further, for Mbone conferencing transmissions Vosaic uses RTP. The decoding formats which are currently implemented are: GIF and JPEG, for images; MPEG-1, NV, CUSEEME and Sun CELLB for video; AIFF and MPEG-1 for audio.

Appendix 2 describes the capabilities of the Vosaic 'plug-in' in more detail. Also included in the Appendix are two extensive essays entitled; 'Real-time Video and Audio in the World Wide Web' and 'Video and Audio: Organisation and Retrieval in the WWW.' These essays explain the approach taken by Vosaic in great detail and give more information about the experiments that have been conducted with the Vosaic/Mosaic browser.

Xing Technology Corporation: <http://www.xingtech.com>

Xing 'StreamWorks' is a plug-in for Netscape Navigator which allows for the viewing of streamed video/audio. This plug-in provides some limited control mechanisms for fast forward, rewind and pause. The video quality is similar to all the products that have been reviewed i.e. the video has poor resolution. The frame rate did not allow for smooth playback and at best this product can deliver a fairly good 'slide show'. For details of products, costs and capabilities from Xing Technology refer to Appendix 2.

Conclusions

The most impressive product was undoubtedly from Progressive Networks. The Real Audio/Video player provided the best quality video at the highest frame rate within the most stylish and functional player. However, it is also fair to say that for this one must pay the highest cost as Progressive Networks' server software is the most expensive.

The most innovative and possibly the most useful product, as regards sign language and the Deaf community, is the Videogram Packager from Alaris. Sending video e-mail allows for better video quality because the process is not time dependent and TCP is used to ensure all the data is received by the client. The video clips tested were running at 15 fps and the resolution was acceptable. A 1651KB video file contained 82 seconds of video. It would be interesting to have further tests on this product to determine how well it handles a video of a deaf person signing.

All the client/server architectures mentioned above do not allow the client to save any of the video files. This is obviously a limitation if one were hoping to store information for future reference.

The conclusion of this chapter is that real-time streamed video is not yet possible. However, great advances have been made in a very short space of time, especially as regards compression techniques and the use of new protocols. The adoption of new time-critical protocols will significantly improve upon present day performance and we can expect a qualitative change in the way video is transmitted within the next few months.

Finally, Appendix 2 contains two essays that dealt with a similar theme to this paper. One is entitled 'Digital Video on the World Wide Web' and the other is entitled 'A/V Streaming brings the Web to Life....Almost.'

Attempts to provide real time streamed video on the Internet are already in danger of being surpassed by the advent of new commercial products which claim to bring real time video conferencing to the Internet. and to the POTS. The next chapter outlines the development of various standards associated with telecommunications and in particular the development of a new standard H.324. This standard has been developed to allow interoperability for products which use very low bit rates for video conferencing, in particular for those applications using the V.34 modem. We believe this is entirely relevant to the situation of the mobile user.

Chapter 3:

Desktop Video-Conferencing

The basic components of a video conference system are:

- Video, requiring a camera and video capture board.
- Audio, requiring microphone and speakers.
- Whiteboard, to show graphs, images, text, documents, or to write on.
- Shared applications.
- Software, to encode and compress the signal and to transmit it to remote sites.

There are many commercial products available¹³ for desktop video-conferencing (DTVC) and choosing a product will depend upon a number of factors. First of all, it will depend upon which network one intends to use and there are a number of options. Video Conferencing may be conducted over the following networks:

- POTS (The Plain Old Telephone System).
- Switched 56
- ISDN (Integrated Services Digital Network)
- LANs (Local Area Networks).
- Internet Conferencing
- Mbone (Multicast Backbone)¹⁴

This research paper is primarily concerned with solutions that are available for video conferencing over networks with low bandwidths, i.e. the POTS and the Internet. In each case, the main obstacle preventing us from having an efficient video conferencing system is the analogue modem¹⁵. Even if the transmitted video is of low resolution and low frame rate, the bit rate necessary for real time video conferencing is still much more than a 28.8 Kbps can handle. To overcome this, compression algorithms are used. Reducing the size of a video file through a compression algorithm, or codec (Compression DECompression) is an integral part of all video conferencing systems. A problem arises however, as different companies may use different codecs which are not always compatible. Interoperability is a key word when it comes to video conferencing and standards are emerging to ensure that the new wave of video conferencing products conform to certain base standards. The International Telecommunications Union (ITU), or rather the Standardisation Sector of the ITU (ITU-T), has developed standards for audio, video, video

¹³<http://www3.ncsu.edu/dox/video/products.html> This is a comprehensive product list with a short review of each product, it is frequently updated. Products that will run on Windows/DOS/OS2 can be found at <http://www3.ncsu.edu/dox/video/features.html#windows>.

Another source for information on videoconferencing products is :

<http://picturephone.com/prodind.htm>. Again, this is a comprehensive list of products with short reviews.

¹⁴ See article entitled 'Mbone, the Multicast Backbone' in Appendix 2. Also included in this appendix is a Frequently Asked Questions paper (FAQ) for a quick summary.

¹⁵ In an article entitled 'Global Video Village', Udo Flohr states; 'There is nothing in the physical infrastructure of the telephone network to make it inadequate for video.' *Byte Magazine* September 96. Pp 138. In this article Flohr looks at Digital Subscriber Lines that use digital modems attached to a standard pair of copper wires.

conferencing and data conferencing, primarily over ISDN. The current standard to which most video conferencing products comply to is H.320.¹⁶

- The ITU-T Recommendation H.320 is entitled “Narrow-Band Visual Telephone Systems and Terminal Equipment.” Narrow-band bit rates are defined to range from 64kbps to 1920kbps. This was designed primarily for ISDN and a 2 BRI ISDN, for example, may use 2 x 64Kbps channels, giving a total bandwidth of 128 Kbps. For more details on this standard refer to appendix 2.¹⁷

Two more recent Recommendations are:

- T.120 This is titled “Transmission Protocols For Multimedia Data.” This recommendation defines multipoint transport of multimedia data. It also enables participants to share data during a conference. This data could be a whiteboard or binary file for example. For more details refer to appendix 2.¹⁸
- H.324 This recommendation is entitled “Multimedia terminal for low bitrate visual telephone services over the GSTN.” This series of recommendations is of direct relevance to this paper as it defines real time data over V.34 modems on the GSTN (Global Standard Telephone Network). In more detail:
 - Video codec H.263: Video coding at rates less than 64kbit/s
 - Audio codec G.723: Speech coder for multimedia telecommunications transmitting at 5.3 or 6.3 kbits/s
 - Control: H.245: Multimedia system control
 - Multiplex:H.223: Multiplexing protocol for low bitrate multimedia terminals.

Using the audio codec leaves either 23.5 kbps or 22.4 kbps for video and overhead.

However G.723 has a silence suppression mode so that audio bandwidth can be used for other data when no audio is being transmitted.

Latest developments in H.324 Compliant Solutions for Video Conferencing.¹⁹

The International Multimedia Teleconferencing Consortium held a ‘virtual’ test session on the 26th Feb. 97. Eleven companies tested their H.324 products and although most were still in the development process some products have recently been released. One company recently announced a H.324 compliant video conferencing unit called ‘ViaTv’²⁰. It is a set top box that fully conforms to the H.324 standards mentioned above. The specification claims are impressive with resolution ranging from 352 x 288 pixels at full CIF, down to 128 x 96 at SQCIF. At the lowest resolution it is claimed that 15 fps is possible. Moving the focus of attention away from computers and on to television receivers is an important step as regards the functionality and ease with which the public can use video conferencing and Internet facilities.

Another company, Lucent technologies²¹ issued a press release announcing the most-cost effective modem chip set that can enable a computer to do video conferencing. The Apollo modem chip uses two integrated circuits, a digital processor chip and a codec chip. It is

¹⁶ Information on ITU-T standards taken from a thesis entitled ‘Desktop Videoconferencing: Technology and Use for Remote Seminar Delivery.’ Written by Leigh Anne Rettinger July 95. The full thesis is included in Appendix 2.

¹⁷ Refer specifically to section 2.4.3.1.

¹⁸ Refer specifically to section 2.4.3.2.

¹⁹ <http://www.imtc.org/main.htm> The International Multimedia Teleconferencing Consortium. ‘The IMTC’s fundamental goal is to bring all organizations involved in the development of multimedia teleconferencing products and services together to help create and promote the adoption of the required standards.’ They are a non-profit organisation and one of the key events that they organise is interoperability test sessions. It is a very informative web site.

²⁰ 8x8 Inc. <http://www.8x8.com/>

²¹ Lucent Technologies <http://www.lucent.com>

fully H.324 compliant and allows for 'digital simultaneous voice and data' (DSVD). The inclusion of a codec chip can significantly reduce the demands made on the CPU whilst improving the performance of the encoding/decoding process. This is all the more important if we consider that mobile computers may not have powerful CPUs and certainly do not have an unlimited power supply.

Another hardware development that will undoubtedly aid the development of video conferencing over POTS is the Intel²² Pentium MMX processor chip. It is claimed that compression/decompression algorithms perform better on an MMX processor chip than alternative makes.

Intel have produced two products that represent the 'cutting edge' in H.324 compliant video conferencing solutions, the 'VideoPhone'²³ (Version 1.2) and the 'Internet VideoPhone' (currently available as a beta test application) Both of these products are based upon the companies 'ProShare' video conferencing technology, originally developed for use over on LANS with an ISDN link. It has now been packaged with the VideoPhone for use over POTS. The application is engineered for OEM platform integration and delivers video/audio compression with a possible frame rate of 12 fps. Many companies, such as Hewlett Packard, Compaq, IBM, Sony and Fujitsu will be packaging this software as part of their PC multimedia systems.

Probably the most exciting development in video conferencing products is the 'Noteworthy Business Video Phone'²⁴ from Toshiba²⁵. This is a complete video conferencing product for use with the Toshiba Tecra /740CDT(as standard)/730XCDT(option) series of mobile computers. It includes a Noteworthy colour analog camera, a CardBus PC Card and video cable connectors, a camera clip for versatile display mount and a complete CD-ROM software package. Furthermore the company is using 'state of the art' 'Zoomed Video' technology (ZV). The ZV architecture allows for high-bandwidth video data to be transferred from a PC Card to the graphics controller without adding traffic to the system bus. This product became available in January of this year and has a suggested retail price of \$499. Toshiba uses ProShare technology (see above) and demonstrated this product in November 1996 using a pentium MMX processor chip.

Conclusion

The combination of hardware and software solutions for video conferencing has for the time being reached its zenith with the Toshiba Tecra 730/740 series of mobile computers.

Problems associated with processing power, efficient codecs, bandwidth limitations and video quality are rapidly being overcome. The Toshiba product is the first of its kind and it is undoubtedly a very important event. The dam, which has been holding back mobile video conferencing, has been broken by Toshiba and we can expect a flood of new products for this market.

The possibilities are considerable and the needs very great. These applications need to be developed further and it would be helpful if experimental work was carried out with these systems but using signing as the language transfer medium.

²² Intel Corporation <http://www.intel.com>

²³ See appendix 2 for further details

²⁴ See appendix 2 for further details

²⁵ <http://computers.toshiba.com>

Chapter 4. Video Compression

Compression Requirements for Sign Language

It is estimated that minimum requirements for sign language interpretation over a video telephone system would be that the video is transmitted at a frame rate of at least 12 frames per second, with a spatial resolution equivalent to the QCIF (quarter common image format) format at 176 pixels per line and 144 lines per frame. If this data were transmitted uncompressed in 24 bit/pixel RGB format, this would be equivalent to a bit rate of approximately 7.3Mbit/s.

This project has looked at the possibilities for transmission over mobile phone networks, which should allow data transmission at bit rates of approximately 28.8kbit/s, and so, even assuming a very high quality channel in terms of error characteristics, and that the entire bandwidth was available for the transmission of the video pictures, there is a requirement for a compression method which can achieve compression ratios of approximately 250 to 1. This would have to be achieved in real time, i.e. 12 frames per second would have to be processed, and the subjective qualities of the reconstructed video would have to be adequate for real time interpretation of sign language.

Current Compression Techniques

Video compression aims to reduce the storage requirements of video sequences, or the bandwidth requirement for their transmission, while retaining sufficient information to reproduce the video sequence, or an approximation to it, which satisfies the requirements of a particular application. This is usually achieved by exploiting the inherent redundancy of the full digital representation of the video frames as arrays of pixels (the spatial redundancy), and the similarities between frames (temporal redundancy). Most compression techniques used for the transmission or storage of video sequences are lossy compression schemes, that is, the reproduced sequence is not, in general, identical to the source sequence. Relaxing the requirement for lossless compression can lead to far more efficiency in representing the video sequence, and the currently popular compression methods control the loss of information in such a way that the differences between the source video sequence and the reproduced video sequence are less noticeable to a viewer, that is, they also exploit perceptual redundancy.

The video compression schemes used in most of the current standards make use

- of motion compensation to predict (the contents of small blocks in) frame contents from previous frames,
- with the facility to encode only differences between predicted frames and the actual frame contents (P (predictive) frame coding),
- along with the motion vectors necessary to reproduce the motion compensation step, and,
- in some schemes, the ability to recreate frames by interpolation of surrounding frames (B (bi-directional) frame coding), in order to exploit the temporal redundancy in the sequence.

The frame is divided into small blocks (typically 8x8 or 16x16 pixels), which are individually compressed, usually using a transform method such as the discrete cosine transform, which exploits spatial redundancy, as well as perceptual redundancy, allowing a more compact representation of the block contents while limiting the noticeable differences between the source contents and the decoded version. Further savings can be made by exploiting similarity between neighbouring blocks, by encoding the components of the transformed representation corresponding to slow variations in value with reference to the corresponding values in previously encoded blocks. Rather than coding an RGB (red green blue) representation of the pixel values, a luminance, and two colour difference or chrominance components are encoded. This allows further savings to be made, and often the chrominance frames are sub-sampled to less than the original resolution, exploiting reduced acuity of the human visual system for colour compared to brightness. Techniques

based on variations of this method have proved very successful in achieving high compression rates in for a wide variety of video applications.

Videotelephony

A method similar to that described above is used in the current standard for video compression in low bit rate videotelephony, which is the ITU H.263 standard. Video compression for videotelephony introduces some factors which must be accounted for in the compression method used. In particular, the frames of the video sequence must be encoded at approximately the same rate as they are displayed, i.e. if it is necessary to display 10 frames per second, the complexity of the method used must be sufficiently low to allow a frame to be encoded in approximately 0.1 seconds, and that the encoder is encoding to a maximum available bit rate, or bandwidth. In the case of H.263, some flexibility is introduced to allow the encoder to vary the way in which the video sequence is handled to account for these factors.

Almost all video codecs (encoder/decoder systems) show a large amount of asymmetry in the processing requirements between the encoding and decoding operations. While decoding is usually fairly straightforward, with the bitstream content determining the operation of the decoder, the degree of flexibility available in encoding the source video sequence means that for the encoder to determine a suitable representation of the encoded video is a much more demanding task. This is reflected in the performance of software based encoders, which often operate at very much lower frame rates than could theoretically be supported by the coding method at the bandwidths available. To use the example of H.263, the encoder has to analyse the video source to determine the most appropriate representation in terms of motion vectors and other factors, before applying the appropriate encoding methods, and this analysis adds a considerable overhead to the total complexity.

Compression for Internet Applications

Although slightly outside the scope of this research report, new compression techniques (Codecs) are being developed to cope with particular problems associated with the provision of streamed video over the internet. In a paper entitled 'Video Compression for the Internet'²⁶, it is stated that a codec needs to fulfil five key requirements if it is to be used for internet applications. These are:

Bandwidth Scalability

This refers to the ability of a codec to deliver compressed video streams over a wide range of bandwidths. On the internet this may range from 20 kbps for a 28.8Kbps modem up to several Mbps for delivery over switched Ethernet and ATM networks.

Resolution, frame-rate, frame-quality scalability

The application should be able to dynamically choose between higher frame rate, improved resolution or individual frame quality. For video with a great deal of motion, higher frame rates are important, for educational videos quality or resolution may be more important to the client.

Fast compression/decompression

The codec should be able to compress video in real-time and should not require expensive hardware solutions, but should be software based. It should not be computationally demanding upon the CPU either at the server end or at the clients end. For video conferencing, the codec will need to be able to allow for multiple decodes and one encode without being too computationally expensive.

²⁶ Appendix A. This essay was taken from the Vxtreme World Wide Web (WWW) Site.
<http://www.vxtreme.com/>

It is provided in order to explain the techniques used by Vxtreme to deliver streamed video over the internet and the authors name is not given. One can only assume that it is a company publication.

Ability to cope with network losses

Individual packet losses must not result in severe degradation of video quality. For the internet, where we can expect packet loss, some form of *redundancy* must be built into the codec. Whilst this codec may not provide us with the highest compression rate files it is necessary if the video is to be of good quality.

Encoding and decoding latency

To achieve greater compression rates it is possible to introduce some latency in to the process. If an MPEG algorithm is used where the B frames (bi-directional) are encoded as a difference from both earlier and later frames, we would need a fairly high number of frames per second (fps) for this latency to be un-noticeable. However, high frame rates are not easily achievable over the internet and at 10 fps waiting for the next few frames will result in a noticeable lag. This type of latency must be optional in any codec designed for the internet.

Of these considerations, resolution, frame-rate and frame quality scalability, fast compression and decompression, and encoding and decoding latency are equally important in point to point applications. Codecs such as MPEG 1 and 2, and standards such as H.261 and H.263 were not developed for use on the internet. MPEG requires expensive hardware support, it is computationally intensive and was designed for quality video broadcasting and CD-Roms. The H.261 codec is similar to MPEG-1 although it is more suitable for video conferencing as it achieves lower latency. However the H.261 codec is not able to deal effectively with network losses and for the internet packet losses can be quite high. Similarly the H.263 codec (the most recent), although designed to deal with the low bit rates associated with 28.8 Kbps modems, copes badly with network losses and results in a significant increase in computational complexity.

New codec standards such as H.323/H.324 (for video conferencing) and MPEG-4 will overcome the problems associated with providing real-time video and video conferencing on the internet to some extent. However, these new standards are open in that they are designed to accommodate new compression algorithms as they are developed and still maintain interoperability. Codecs are now seen in a more modular way i.e. modules to be slotted in or taken out of an application as needs and requirements change.

In appendix 2²⁷ there is an extract from a thesis written by Carl Johan Berglund which deals specifically with compression techniques for the internet.

Appendix 2 also includes another essay taken from the Vxtreme WWW site entitled 'Enabling Interactive Video Over the Internet'.²⁸

Second Generation Video Compression Techniques:

Over the last few years, it has become apparent that prospects for significant improvements in the compression ratios and reconstruction characteristics of techniques using hybrid DPCM/Motion Estimation/DCT coding algorithms are limited. Although research into new variations on these techniques continues, and further improvements are still being made, a lot of the current research work in video compression is based around developing new methodologies and coding schemes, which will be needed to provide the very high compression ratios demanded by new applications. Most of these techniques are still at a very early stage in their development, and few have found their way into available video coding applications yet, but the level of research in these areas, together with some apparently promising results, suggests that significant progress may be made in these areas in the next few years.

²⁷ 'Digital Video on the World Wide Web' Masters Thesis in Computer Science at Kungl. Tekniska Hogskolan. Sdtockholm Switzerland. Written by Carl Johan Berglund, 5th November 1996. Ch 2 'Video Compression Methods'.

²⁸ Taken from <http://www.vxtreme.com> Author unknown.

The most promising of these techniques, known as "second generation schemes" are based mainly on a high level description of the image data in terms of visual primitives. Major projects to look into these schemes, recognising their paramount importance to applications such as very low bit rate mobile video communication, have been proposed by ACTS (TASK AC103:Advanced Second Generation Image Coding Schemes) and suggested examination of a range of techniques including advanced segmentation and motion estimation, region based coding, object oriented coding, 3-D model based coding, multi-scale coding, morphological coding, and hybrid waveform/object coding. Due to the large amount of research work being carried out in these areas, and their rapid development, it was not possible to review all of the available information of developments in these areas, but some information and recent results are outlined below, along with summaries of some of the basic methods being used, and current research projects

Model Based Coding

Most of these methods are based, at some level, around an analysis/synthesis approach. That is, rather than using a transform based method as in current methods, the coding method attempts to identify the objects depicted in the sequence, and rather than transmit a compact representation of the actual data that makes up the picture, a description of the scene contents in terms of parameters of some model of the objects represented, or at least a higher level description of their shape and texture, is encoded.

Model based coding techniques are a major research area, and could provide methods for producing video telephones operating at much lower bit rates than any current system. Methods currently being investigated are focused on "head and shoulders" type video, and are being considered in MPEG4, to utilise the SNHC (synthetic and natural hybrid coding) principle for video coding, and proposals have been invited for general head and face modelling techniques.

The methods being considered involve analysis of the image sequence to determine the most appropriate parameters to represent their content in terms of a three dimensional model of a particular person's head and face, and transmitting only these parameters, along with possibly some additional information for accurate reproduction of textures/ etc. The decoder can then generate an artificial picture with sufficient detail to allow facial characteristics etc. to be determined. Current systems have had some success with this approach, but this is generally seen as a futuristic approach, which has the long term potential for videotelephony at very low bit rates (it is estimated that 1-2kb/s would be sufficient).

The model based approach has several advantages over traditional coding methods for videotelephone applications. One factor is that it does not necessarily attempt to reproduce accurately the source video sequence in all respects (although this can be seen as a long term aim); the method is also appropriate in situations where it is necessary to reproduce only those aspects which are considered essential for communication, such as facial expressions and, in the case of sign language, posture and hand/arm movement.

Another factor is that, whereas traditional coding schemes take a single video sequence as the only input to the coding scheme, and attempt to reproduce it, the change in emphasis in model based coding allows the encoding system to take full advantage of the actual presence of the subject to aid the encoding process. For example, the ability to analyse facial features to determine the model parameters is greatly simplified by the use of stereo cameras. This provides additional information which is not available from the video sequence produced by a single camera, and allows much more reliable analysis by comparison of the stereo images.

This method is used in several techniques currently in development, and extension from head and shoulders video to upper body motion capturing hand movement should be possible. Although hand movements are less limited in terms of screen position than facial movements, they are less complex and the necessary accuracy of reproduction is likely to be

much less, with problems such as those encountered when trying to reproduce convincing facial expression being less apparent.

Frame Segmentation

Improved segmentation of the frames of the image sequence can also lead to significant improvements in compression, and segmentation techniques are another area where progress is currently being made in image compression.

This may be particularly appropriate to sign language applications, as different regions may have different requirements in the reproduced video. For example, fast hand movements should be reproduced, and so it may be necessary to encode the corresponding regions of the sequence at higher frame rates, facial details may require a higher resolution, and the background region may be discarded without hindering sign language communications. Although it is usually quite complex to segment accurately an image sequence into regions corresponding to particular features, or with particular characteristics, this may again be facilitated by techniques such as stereo cameras. In fact, segmentation of the subject from the background is performed as an initial step in model based coding, and the complexity of this process using stereo images is relatively low. One potential problem in mixing resolutions and frame rates within a single image sequence is that the artefacts of recombining these features may be distracting. This would have to be considered in details if this were proposed as a serious method for compression of sign language video.

Advanced encoders using new techniques

Most full video compression schemes currently being developed use a hybrid approach, combining features of one or more second generation coding techniques with techniques from current coding methods. One example is dynamic or competitive coding, in which several coding techniques are attempted in each region of an image identified by a segmentation procedure, and the most efficient for the particular region is used. This can lead to better compression, although at the expense of computational complexity. Some compression techniques based on this type of approach have proved quite successful, and seem to approach the quality of reproduction required for sign language at the approximate bit rates being considered in these applications. All of the possible coding methods listed in the ACTS task description appear to have some capacity to improve coding efficiency compared with current standards.

Examples of projects and papers investigating new coding techniques are the RACE2 MORPHECO project, which investigated morphological techniques, RACE MAVT (mobile audio visual terminal), which considered various schemes as improvements to H.263, ACTS MoMuSys, which continues the MAVT project.

Promising results

Morphological coding also has some potential for improving compression. New proposals have recently been introduced by Nokia, using segmentation and advanced motion compensation along with DCT transform based coding and vector quantisation, and work is underway in collaboration between Orange and the University of Strathclyde, on a method using several new techniques, as well as neural network based analysis. The initial results of both of these projects look very promising.

Chapter 5: Possibilities and Specifications for video

In the previous chapter, information was given on the current compression methods, and methods in development which could lead to improvements in the next few years. Video compression is a rapidly developing field, and this is reflected in the current activities in defining new related standards. The requirements and specified aims of these standards give some indication as to the facilities and improvements which should soon be available. Currently, much of this work is focused on the definition of the new MPEG4 standard, and the related H.263L standard. Both standards are due to be completed in November 1997.

MPEG4

One of the main objectives for this project was to examine the coding capabilities of MPEG4 to assess its ability to code video with suitable properties for sign language transmission at mobile phone bandwidths. This objective is hindered by the fact that the MPEG4 standard is still in development, and also by the envisaged flexibility of the MPEG4 standard.

MPEG4 was originally intended as a standard for low bit rate video coding to enable videotelephone applications over a standard telephone line, and other applications requiring high compression ratios for video, to be produced. With the introduction of the ITU-T H.263 standard, the focus of MPEG4 development changed, as H.263 was already capable of delivering a reasonable quality of video at these bit rates, and it was considered unlikely that significant improvement could be achieved in the time scale set for the production of the MPEG4 standard. MPEG4 currently addresses the need to establish a

"universal, efficient coding of different forms of audio-visual data (audio visual objects). These objects can be of natural or synthetic origin".

The approach taken in reaching this goal is based on the definition of a set of coding tools for audio-visual objects capable of providing support to different functionalities, such as object based interactivity and scalability, and error robustness, in addition to efficient compression, and also a syntactic description of coded audio visual objects, providing a formal method for describing the coded representation of these audio-visual objects and the methods used to decode them. The MPEG4 standard is due to be finalised in November 1998, but a lot of issues still need to be clarified.

The video output of an MPEG4 decoder will be formed from the composition of any number of individually coded objects, called "visual object planes" (VOP s). These can be thought of as the contents of arbitrarily shaped regions of the final frame. The standard will provide mechanisms for multiplexing these objects into one data stream, mechanisms to guarantee particular properties of that stream, such as error recovery, and conveying information to the decoder describing how the object should be decoded and composited in the final video output.

The objects encoded are intended to some degree to represent individual objects of importance to the application, and by their separate encoding, more information is available to the decoding application concerning, for example, the positions and properties of important objects on the screen, so that the degree of object based interactivity anticipated in distributed multimedia applications in the near future can be more easily supported. The objects can represent either features in a particular video sequence, an entire video sequence, or artificially generated objects which must be fully rendered by the decoding application. The ability to encode artificially generated objects as well as traditional image sequences, and to composite these into the video output, is known as "synthetic and natural hybrid coding" (SNHC), and is a major new feature in MPEG4 which was not addressed by previous coding standards.

In defining the MPEG4 standard, the following requirements were made for the capabilities to be addressed:

1. Content-based interactivity

Content-based multimedia data access tools

MPEG4 shall provide efficient data access and organisation based on the AV content. Access tools may be indexing, hyperlinking, querying, browsing, uploading, downloading and deleting.

Content-based manipulation and bit-stream editing

MPEG4 shall provide a syntax and coding schemes to support content-based manipulation and bitstream editing without the need for transcoding. This means that the user should be able to select one specific object in the scene and perhaps change some of its characteristics.

Hybrid natural and synthetic data coding

MPEG4 shall support efficient methods for combining synthetic scenes with natural scenes. This functionality offers something new to the image world: the harmonious integration of natural and synthetic AV objects. This represent a first step towards the unification/integration of all kinds of AV information.

Improved temporal random access

MPEG4 shall provide efficient methods to randomly access, within a limited time and with fine resolution, parts from an AV sequence.

2. Compression

Improved coding efficiency

MPEG4 has set as its target to provide subjectively better AV quality compared to existing or other emerging standards, at comparable bit rates.

Coding of multiple concurrent data streams

MPEG-4 shall provide the ability to efficiently code multiple views/soundtracks of a scene as well as sufficient synchronisation between the resulting elementary streams. For stereoscopic and multi-view video applications, MPEG-4 shall include the ability to exploit redundancy in multiple views of the same scene, also permitting solutions that allow compatibility with normal video. This functionality should provide efficient representations of 3D natural 'objects' provided a sufficient number of views is available. Applications such as virtual reality can substantially benefit from this capability.

3. Universal access

Robustness in error-prone environments

MPEG4 shall provide an error robustness capability. Particularly, sufficient error robustness shall be provided for low bit rate applications under severe error conditions. Note that MPEG4 will be the first AV representation standard where channel characteristics are considered in the specification of the representation methods.

Content-based scalability

MPEG4 shall provide the ability to achieve scalability with a fine granularity in content, spatial resolution, temporal resolution, quality and complexity. Content-scalability may imply the existence of a prioritisation of the objects in the scene. The combination of more than one scalability case can yield interesting scene representations, where more important objects are represented with higher spatial and/or temporal resolutions.

MPEG4 shall also provide subjectively better AV quality at comparable bit rates compared to existing or other emerging standards.

In the context of sign language transmission in over limited bandwidths, several of the envisaged new features of the MPEG4 standard may be advantageous. The facility for

encoding a video sequence as a collection of arbitrarily shaped regions, which can be encoded separately may allow for better control over the degradation characteristics of different regions of the image, as well as allowing the enhancement of regions of particular importance.

The above discussion indicates that, rather than providing a general encoding method which provides limited flexibility to the encoder, MPEG4 intends to be capable of addressing a wide variety of different applications, and allow a much greater degree of flexibility than allowed by previous coding standards. The existing coding methods used in, for example, the H.263 standard will be supported in MPEG4, which will also incorporate newer methods. In the short term, existing systems give some indications of the minimum standards that can be expected, although the greater degree of flexibility in MPEG4, the ability to achieve more graceful degradation in limited bandwidth situations, and the ability to particular regions in the image frames, should allow for some improvement. Investigation of newer techniques, related emerging standards and techniques for improving the readability of decoded sign language video give a better indication of longer term trends.

H.263L

The main focus of MPEG4 activity is in providing suitable coding mechanisms for new multimedia applications, and, although improved coding efficiency and quality is still a major part of this work, the provision of very low bit rate video coding methods is no longer the main priority. Work on video coding is being carried out concurrently by the ITU H.263L Advanced Video Coding Ad-Hoc Group, in collaboration with MPEG4, in developing the next generation of low bit rate video coding standards. The short term goal is the definition of the H.263+ standard, which is an extension to the current H.263 standard, and a new H.263L standard, for which a wider range of coding strategies is under consideration.

This project was proposed in April 1996, and H.263L is intended to be completed in November 1998, coinciding with the standardisation of MPEG4, and its proposals are expected to be incorporated within the MPEG4 standard as tools for very low bit rate video coding. The work is targeted at "real time audio/visual conversational services" (video telephony) operating over channels of less than 64kbps capacity. The adopted technology should be able to provide enhanced error robustness in order to accommodate mobile networks, will address video delay and codec complexity.

New criteria are given for video quality characteristics in terms of their suitability for particular applications, including face recognition, emotion recognition, lip reading, sign language reading, object recognition, and text reading. These are assigned different levels of priority, but amongst these, a quality of reproduced video sufficient to allow sign language reading is considered to be essential for any proposed coding scheme (although no information is available on how they will be determining the suitability of schemes for this application). In addition, regular delay must be less than 250ms, and should be less than 150ms, and QCIF resolution of luminance images must be supported (full details are in Document LBC-96-136, Appendix 1).

This implies that any coding methods which fulfil the stated requirements will necessarily fulfil all of the requirements of a sign language mobile videotelephone system so far identified, with the exception of strict bandwidth requirements (although the work seems to be concentrated at bandwidths approximating those we are considering) and processing/memory requirements for the encoder and decoder, the specifications of which have not so far been determined, but are factors being considered.

Submissions for H.263L

The proposed "open submission phase" in the development of this standard, in which proposals for techniques or technologies suitable for the needs of the project will be received, is due to end in June 1997. Proposals have already been received (Feb97) for

some aspects, including a full H.263L proposal from Nokia, which apparently already meets the requirements for video encoding at low bit rates suitable for sign language interpretation, and achieves approximately 3dB PSNR improvement over the current H.263 standard at bit rates suitable for videotelephony over PSTN, or up to 55% reduction in bit rate at approximately equal quality. This proposal (reproduced in Appendix 1 document LBC-97-029) is based around partial segmentation of the image sequence, and an improved motion estimation model, and follows an initial proposal in July 1996. A presentation on the same method was also given at the ACTS Mobile Telecommunications Summit in November 1996 (reproduced in Appendix 1 "Advanced Video Coding Scheme for Low Bit Rate Applications"), where it was considered as a possible video coding method for use in the ACTS Moments project.

This method seems very promising in terms of bit-rates required and quality of reproduced video; the quality of video possible over a 28.8kb/s channel should be roughly similar to that produced by an H.261 based system over a 128kb/s ISDN link, which we have seen allows for a reasonable amount of sign language communication. The processing requirements for this method may cause problems, particularly in a mobile solution; decoding in real time is possible using current hardware, but encoding is not (the November 1996 paper quotes an encoding time of at least 20 seconds per frame on a HP9000 computer, although this may have been improved in the subsequent proposal). The authors claim that this may be overcome with the more powerful hardware which will soon be available, and also give an indication that considerable savings in computational cost can be made at the expense of a small increase in bit-rate.

The following tables compare the performance of the proposed method with H.263 using a variety of standard test sequences:

Other techniques which have been considered for H.263L include a robust object based video coding method from TI, and a new compression transform from the University of Strathclyde. The University of Strathclyde have been working in collaboration with Orange to produce a new low bit rate coding method, which utilises neural networks, image segmentation, transform based coding, motion estimation and block substitution. Many aspects of the operation of this method remain confidential, and development of the scheme is still in progress, but initial indications are promising, and demonstrations of early versions of the coder have shown one way video transmission over a single 9.6kb/s mobile channel, with real time encoding, and significant improvements over H.263 have also been achieved. In addition to work on video compression, the operation of other factors in the transmission of video is also being considered in the development of H.263L. These include a MultiLink Protocol for, characteristics necessary for archiving and retrieving files of compressed video, and provision of reliable communications over mobile channels.

Chapter 6 - Trials

Performance of H.263 at low bit-rates:

Several H.263 systems were evaluated at maximum bit rates of 28.8kb/s to investigate whether this standard could deliver video quality sufficient for sign language communications.

Initial tests were performed using a software implementation of this scheme (Cybersoft VisualEyes), using Quickcam video cameras. This implementation encodes and decodes in real time on a standard PC (tests used a 166MHz Intel Pentium machine and a 100MHz AMD Pentium machine), but the frame rates possible using this system proved too low for sign language communication to be possible. The system was capable of encoding at best 1-2 frames per second, and the system is better suited to sending a series of still images (which may be appropriate to some video-conferencing applications) than for anything approaching smooth motion. This system did not give a fair indication of the full

capabilities of H.263 coding, although it did give an indication of the quality which is achieved in software implementations.

A prototype system under development by Motion Media plc was also evaluated. This system used more specialised hardware, and made more extensive use of image processing techniques to improve the quality of the reproduced images, and was also able to make better use of the techniques available in the H.263 standard. The Motion Media system was more flexible, and frame rates were achieved in a simulation of a 25kbit/s bit rate which allowed movement to be followed, fairly easily, but this was at the expense of a lot of spatial detail, and interpretation of sign language, lip movements and facial expressions using this system, whilst possible to a limited degree would still appear to be problematic.

Demonstrations of test sequences encoded using H.263 which were available on the internet were also examined. These gave an accurate demonstration of what can be achieved using this standard, and the decoded sequences, while showing considerable levels of "block" artefacts and poor spatial resolution when the sequence contained a lot of motion, still indicated that some sign language communication may be possible at bit rates of 20kbit/s provided that real time encoding of a similar quality could be achieved by a real time encoder.

The H.263 standard is considered by some people to be approaching the limits of what can be achieved using hybrid DCT/motion estimated predictive coding techniques at very low bit rates, and is becoming widely used for video conferencing applications over PSTN. It is clear that the possibilities for using this method for sign language communication are limited, particularly in software implementations running on standard hardware. No software based encoders we came across were capable of high quality H.263 encoding in real time at acceptable frame rates, although coders making use of specialised hardware, or those utilising MMX technology to improve the efficiency with which video data can be processed, should show significant improvements. In general, H.263 could be expected to achieve video compression suitable for a reasonable level of sign language interpretation provided that at least 20kb/s was available for transmission of the video signal, in situations where the total amount of motion depicted in the scene is limited.

Preliminary Findings

The results of MPEG4 core experiments into low bit rate video coding showed that the basic coding methods used by these standards still gave the best combination of coding efficiency and quality of reconstruction, and so the coding methods were retained as MPEG4 tools.

The envisaged MPEG4 method for very low bit rate coding can be seen as a superset of the methods employed by H.263, H.261, MPEG1, MPEG2, and may include new coding methods when techniques leading to significant improvements become available.

The capabilities of current coding methods therefore give some indications of the likely performance of MPEG4 based codecs. H.261 operates at bit rates of $px64$ kbit/s (i.e integer multiples of this rate), and is predominantly used for videotelephone applications over ISDN phone lines. Tests were conducted using a PictureTel 100 system at 128kbit/s, using a coding method compliant with this standard (described below).

Informal Tests

Tests showed that a reasonable quality of video was possible using this system, and was sufficient for sign language communication in certain circumstances. The approximate frame rate of the reproduced images was approximately that which we consider necessary for sign language communication, although some difficulty was experienced in interpretation when fast hand movements were used. Edges of moving regions in the image showed considerable degradation, hindering the ability to follow hand movements accurately. The frame rate for regions of the image depicting rapid movement was slow, and important frames were often "dropped". A delay of about half a second was introduced in

the transmission of video data, making natural conversation using both sign and voice difficult. If the scene depicted contained a lot of motion (for example with background movement or movement by the subject), the reduction in frame rate and the block artefacts introduced reached unacceptable levels. Improvements introduced in the H.263 standard can give a 50% reduction in bit rate over H.261, and new algorithms are being produced which can deliver a further 50% bit rate reduction over H.263, and so the capabilities of this system may give a fair indication of the quality of video which will be possible over 28.8 kb/s channels in when MPEG4 is introduced.

Sign Language and Video Conferencing Trials

One aim of this project was to evaluate the use of sign language in a video conferencing system.²⁹ Our primary interest related to the problems of using sign language in a two dimensional medium as opposed to real life. Furthermore, we were interested in problems that might arise given the linguistic structure of sign language.

The test required four sessions and was conducted over a period of two days. Four subjects were involved in the testing and their sign language abilities ranged from 'fluent' BSL (British Sign Language) to 'fluent' SSE (Sign Supported English - mixed BSL and English).

Test 1

This test involved the fingerspelling of twenty-four words. The words chosen generally consisted of letters that we expected to be more difficult to recognise. The following words were used:

Group One

Men	Lone	Acorn	Liners
Run	Jeep	Jerry	Arming
Vee	Mean	Roman	Vanity
Mat	Lank	Align	Lurked
Rot	Jobs	Shave	Bright
Van	Musk	Story	Asleep

Group Two

Man	Live	Alert	Learnt
Rim	Joke	Party	Arrays
Via	Moat	Liven	Cavity
Jog	Easy	Money	Victor
Rub	More	Honey	Rumour
Low	Core	Blown	Lonely

As can be seen six 'three', 'four', 'five' and 'six' letter words were used. The letters we expected to cause most difficulties were:

L M N R V ... and to a lesser extent: E I and O .

The tests were conducted as follows.

Each student, in turn, were asked to watch a member of staff fingerspell one of the group of words above. This test was conducted without the videoconferencing system.

²⁹ The system we used was the PictureTel 100 Video Conferencing product. It was run over an 2 BRI ISDN direct link, which gave us a bandwidth of 128Kbps.

Students A and B were asked to write down the sampled words from group one. Student C and D were asked to write down the sampled words from group two. Students were permitted to ask the word to be re-spelled as many times as they wished. The time taken for this part of the test to be completed was recorded.

Student A Time taken = 2 minutes 48 seconds
Student B Time taken = 3 minutes 19 seconds
Student C Time taken = 3 minutes 29 seconds
Student D Time taken = 3 minutes 45 seconds

We were not concerned with the students understanding of the sign, or if they eventually made one or two errors when the *word* was written down . In these tests only two words were recorded wrongly, one student recorded **Livers** instead of **Liners**, and another **Line** instead of **Live**. In both cases confusion resulted from the student mistaking the V and N letters.

The tests were repeated and students were asked to fingerspell the words to each other using the PictureTel videoconferencing system.

Student A fingerspelled the group 1 words to student C
 Student C fingerspelled the group 2 words to student A
 Student B fingerspelled the group 1 words to student D
 Student D fingerspelled the group 2 words to student B

Again, the time taken for this test to be completed was recorded. Students could ask for the word to be re-spelt any number of times.

The results were as follows:

Student A Time taken = 3 minutes 56 seconds
Student B Time taken = 6 minutes
Student C Time taken = 6 minutes 35 seconds
Student D Time taken = 8 minutes 25 seconds

Conclusions

Results	Face-to-Face	Using PictureTel	Difference
Student A	2 Min 48 sec	3 Min 56 sec	1 Min 8 sec
Student B	3 Min 19	6 Min	2 Min 41 sec
Student C	3 Min 29	6 Min 35 sec	3 Min 6 sec
Student D	3 Min 45	8 Min 25 sec	4 Min 40 sec

In all cases, reception of the fingerspelled words took longer through PictureTel, in one case significantly longer. In fact this was an interesting result in that the amount of difficulty presented by the videoconferencing system to the student seemed to be in direct relation to the students signing skills. Student A was the most proficient user of BSL , students B and C were very capable BSL users, but student D used a great deal of SSE (Sign Supported English). Whereas student Ds time for the face-to-face test was similar to the other students (although it was also the slowest) the time taken for student D to complete the test was far longer than the other students. In fact this confirms earlier trials which were conducted with two Centre for Deaf Studies staff members who are ‘native’ BSL users.³⁰

³⁰ Before we conducted trials with students we ran some preliminary tests using two members of staff. The staff members were fluent in BSL and were native users of the language. It was from these tests that we developed both the words to be used for fingerspelling and the words to be used for the individual signs.

As expected students found some letters more difficult to recognise than others. In this test it was clear that students reception skills were put to the test when using the videoconferencing system. Students reported that a greater level of concentration was needed and even then students were forced to ask for the words to be re-spelt quite a number of times. All students found the letters **L M N** and **R** the most difficult to recognise and words including these letters were far more likely to be misunderstood.

Test 2

This test was similar to the first test but instead of words being fingerspelled they were signed. The design of the test was exactly the same as test 1. Two groups of words were chosen and again they were not chosen at random. The words in column one are signed on or around the face, in column two on or around the body and in column three the signs were 'open'. The purpose of this was to ascertain whether or not recognition of signs related to their spatial position.

Group One

Useful	Really	Wide
Quiet	Insurance	Conference
Sly	Calm	Argue
Scar	Like	Sun
Moustache	Mine	Rain
Pretend	Operation	Snow

Group Two

Shout	Switzerland	Thunder
Sick	Farm	Encourage
Shame	Garden	Train
Clever	Angry	Clouds
Glasses	Frustrated	Expensive
Pink	Confidence	Cheap

Each student, in turn, was asked to watch a member of staff sign one of the group of words above. This test was conducted without the videoconferencing system.

Students A and B were asked to write down the words from group one.

Student C and D were asked to write down the words from group two.

Students were permitted to ask the word to be re-signed as many times as they wished. The time taken for this part of the test to be completed was recorded.

Student A Time taken = 30 seconds

Student B Time taken = 1 minutes 46 seconds

Student C Time taken = 1 minutes 40 seconds

Student D Time taken = 2 minutes 20 seconds

The tests were repeated and students were asked to sign the words to each other using PictureTel.

Student A signed the group 1 words to student C

Student C signed the group 2 words to student A

Student B signed the group 1 words to student D

Student D signed the group 2 words to student B

Again, the time taken for students to receive the information and write it down was recorded. Students could ask for the word to be re-signed as many times as it was needed.

Student A Time taken = 2 minutes
Student B Time taken = 2 minutes 53 seconds
Student C Time taken = 3 minutes 32 seconds
Student D Time taken = 3 minutes 48 seconds

Results	Face-to-Face	Using PictureTel	Difference
Student A	30 sec	2 Min	1 Min 30 sec
Student B	1 Min 46	2 Min 53 sec	1 Min 7 sec
Student C	1 Min 40	3 Min 32 sec	1 Min 52 sec
Student D	2 Min 20	3 Min 48 sec	1 Min 28 sec

Conclusions

In contrast to the fingerspelling test, the difference between face-to-face communication and communication using PictureTel was not dramatic. This is especially true for student D who managed to complete the task in a similar time to student C. It was clear from this test that understanding signs is far easier than understanding fingerspelling through a videoconferencing system. Some points to consider are:

- Students sometimes failed to understand a sign because it was difficult to know the context in which the particular sign was used. For this, it is necessary to be able to recognise the mouth pattern that accompanies the sign. All four students reported that mouth patterns were not always discernible.
- One or two signs were signed using a regional version of the sign. When this occurred the student informed the observer that although they could recognise what was being signed they did not know its meaning.
- The spatial position of the sign had no effect on the students ability to recognise it.

Test 3

This test was concerned with the students ability to follow a short story or joke. The story or joke was signed as follows:

Student A signed a story to Student B

Student B signed a joke to Student A

Student C signed a story to Student D

Student D signed a joke to Student E

The results of this test were not timed but students were asked to tell us their understanding of the joke or story and to repeat as much as possible of it to the observer. The results were a little surprising as all four students followed the joke/story with very little trouble. All were able to faithfully reproduce the story or joke to the observer and the two jokes managed to get a laugh! This is not an entirely humorous statement as it showed that students were able to watch the screen in a relaxed manner and enjoy the information that they were receiving. This despite the fact that using a videoconferencing system does require more concentration. In fact, the main conclusion of this part of the test is that video conferencing is suitable for sign language communication and that the problems inherent in it can be overcome. Students in this part of the test seemed to have benefited from having participated in the earlier tests, in short students seemed to have learned how to use PictureTel. It is probably accurate to say that the more the system was used the easier it became. This may relate to

the psychology of human-computer interaction more than it does to a particular mode of communication.

Test 4

This test involved letting the students converse informally. We wanted to observe how the system would be used in a very informal situation and whether the students would enjoy communicating via this channel. Each pair of students conversed for a total of ten minutes after which we asked some questions and asked for comments. The questions we asked were:

1. Would you use videoconferencing in a social way if it was available?
2. Did you find it more tiring ?
3. What are the main problems associated with signing over a videoconferencing system?

The answers we received were very positive. All four students thoroughly enjoyed using PictureTel and answered in the affirmative to question 1.

Surprisingly in answer to question two, all four students said that they did not find it more tiring when communicating via this channel. This was surprising because our earlier tests showed that conveying the same information did take significantly longer when using videoconferencing.

The main problems that students identified were:

- If the camera picked up any movement behind the signer it was very distracting.
 - It was very difficult to communicate with more than one person at a time. A short test involving three students at the same time proved to be very confusing and resulted in very stilted conversation.
 - Fingerspelling proved to be the major problem.

Discussions

These tests indicate the viability of personal communication in this way. However, the mobile development is not yet able to deliver video. There are relatively few applications available and most allow only viewing.

Chapter 7 - The way forward and specifications

Requirements and Advantages in Sign Communication

From our tests with deaf people in the trials of video telephones using ISDN lines, it is clear that there is a high level of enthusiasm for this type of service. Reasonably cheap home video-telephone systems, in the form of TV set top boxes and PC hardware add-ons will become widely available later this year. These could go some way towards improving the usability of telecommunications for the deaf community.

To review the requirements for mobile sign language communication in a mobile system, the video should be compressed as much as possible, but ideally this should be sufficient to be transmitted at bit-rates not exceeding 28.8 kbit/s. This has to be accomplished over mobile channels, and so enhanced error robustness must be provided. Encoding and transmission must be possible in real time at a rate of over 10 frames per second. Most importantly, the quality of reproduced video must be sufficient for sign language interpretation.

Current Capabilities

Compression methods currently in use for videotelephony applications may go some way towards allowing a limited amount of sign language communication over more limited bandwidths, although the solutions offered are less than ideal, and fall some way short of meeting all of the requirements. H.263 systems are currently being introduced which make use of either specialised hardware or the MMX extensions for PC processors which should allow for improved real time encoding using this standard. As these systems have only recently become available, it was not possible in the course of this preliminary investigation to fully assess the potential they offer.

In terms of current technology, the requirements for a mobile system which would enable communication by sign language over mobile channels are extremely demanding. While the new methods seem to address problems of bandwidth, robustness and quality of reproduction, this is at the expense of processor requirements if real time encoding is to be possible.

A demonstration mobile multimedia terminal is being developed in the ACTS MoMuSys project, and this gives a fair indication of the requirements for a mobile system for sign language communication making use of compression techniques suitable for general video coding. In this project, a portable PC was used, with added DSP cards to enable real time audio and visual processing for video encoding, although it is noted that this does not meet the criteria of lightness and compactness which would be ideal for a mobile terminal. This was seen as an acceptable compromise for a demonstration terminal, particularly as there was little point in developing the specialised hardware which was considered necessary while relevant standards were insufficiently defined. In the MoMuSys project, higher bandwidth mobile channels (64kbit/s) were considered, as these are likely to be available in the near future. It is noted, however, that such services are likely to be expensive, and so efficient coding is still extremely important.

The video compression in the MoMuSys system was implemented on a TMS320C80 processor, described as a "powerful chip containing a master and four slave processors" and intended specifically for DSP (digital signal processing). This proved sufficient for the coding method used, but no figures are available as to the quality of coding achieved, and even this may be inadequate for the more advanced coding methods which may be necessary.

The requirements of sign language in video telephony are in many ways far more demanding than those of more conventional use of this technology. While the current standards for videotelephony allow for some level of communication at the very low bit rates which are being considered here, the solutions offered are patchy, and the quality of reproduction has noticeable deficiencies even when dealing with a conventional "head and

shoulders" type of content characteristic of conventional videoconferencing. In particular, current software implementations of this standard do not approach the quality required for this application.

In traditional applications, the quality produced by software coders may be acceptable, as the amount of information expected to be conveyed by the video picture is minimal, but in attempting to use this technology for sign language communication, even at less restricted bandwidths, problems arise both from the fact that the transmitted video contains a lot more motion than in head and shoulders video, which is more problematic for video compression, and also due to the additional requirements in the quality of the video reproduced, which must accurately reproduce fast hand movements and facial expressions.

Examples of what could in theory be achieved using the current standards suggest that a bandwidth of 28.8 kbit/s is sufficient for a limited amount of sign language communication, but that the degradation characteristics are unhelpful. Real time encoding capabilities are limited, and the ability to make full use of the potential for fairly high quality video requires encoding at a frame rate which is not achievable on inexpensive, standard hardware in software implementations, and the processing requirements may be particularly prohibitive in a mobile environment, particularly when factors such as battery life are considered.

The field of low bit rate video coding is a very active research area, in which significant advances are being made. New proposals for H.263L have been submitted in the last few weeks, and these appear to be very close to reaching, and may even surpass, the minimum requirements in terms of compression ratios and quality of the reproduced sequence that would be required for sign language communication, although the processing requirements for the recent Nokia proposal would appear to be prohibitive for a mobile implementation on current hardware. Providing these methods can be incorporated into flexible coding schemes, as is consistent with the MPEG4 approach, tailoring the schemes to the specific requirements of a mobile sign language telephone, and making use of additional assumptions and information available for this particular application may facilitate significant reductions in the complexity of the coding process and alleviate possible problems. Other methods being considered for H.263L appear to be less complex, and so may be more suitable, but these systems are still in development, and so no definite information on their characteristics is available.

In general, in terms of video compression requirements, a reasonable level of sign language communication looks likely to be possible using the techniques available in new standards, the definitions of which are due to be finalised in November 1998, at bandwidths which should soon be available over mobile communication channels. The full implementation of a mobile videophone system relies on new developments in various fields (mobile communications, video compression, image processing, enhanced multimedia processors, robust transmission over mobile channels), and capabilities may be enhanced by techniques which would not usually be considered in a traditional mobile phone application (background segmentation, feature enhancement, contrast enhancement, model based enhancement of particular features).

Implications

There are obvious practical difficulties in using a video telephone, particularly for sign language, which differ from those of mobile telephones. It would obviously be very difficult to use the system in many situations where an ordinary mobile phone could be used by a hearing person, due to the need to place the system at a certain distance from the subject (to record), leaving both hands free for signing. This has some advantages, such as the likelihood that an external power source will be available, but introduces new considerations as well.

The more limited range of situations in which a mobile videotelephone system could be used compared with an ordinary mobile phone means that the ability to use "sign-mail" may be

extremely important. This is not so demanding in terms of compression compared with real time videotelephone applications, and could probably be adequately implemented using existing methods. In particular, there is no requirement for *real time* compression, and the bandwidth restrictions can be relaxed. New capabilities are being introduced into mobile networks which would assist the practical implementation of this type of service, for example, the capability to utilise available bandwidth but at a low priority should be available.

As mentioned previously, the minimum requirement in terms of resolution and frame rate for sign language to be practical is a resolution of about 176x144 pixels, a frame rate of about 12 frames per second, and a maximum delay of less than one second. These requirements are all addressed in the H.263L proposals, and so, if these requirements are achieved, this standard should provide a good basis for a sign language telephone system. The bandwidth limit of around 28.8kb/s seems achievable in the proposals being considered. Portability is obviously a major consideration, and the processing requirements of the proposed system would appear to preclude any implementation in the immediate future.

Future Needs

Further tailoring of the coding methods used in standard systems to make the characteristics more suitable for sign communications may be possible. Several important factors have been identified for sign language interpretation of a video image, and more work needs to be done to clarify these, and to find out to what extent the compression characteristics in relation to these can be improved within the current and forthcoming standards. Artificial enhancement of aspects of the image sequence may also be helpful in improving sign communication. These could merely be improved segmentation and possibly contrast improvement, or could extend to more complex operations, such as model based enhancement of particular features. This method has been used in the ACTS VIDAS project, to enhance the representation of lip movement by analysing the received speech and compositing an artificially generated mouth image onto the received video. The application of this particular method in communications between deaf people are obviously limited, as it relies on the speech of one participant, but it is expected to facilitate lip reading in communication between a deaf and a hearing person.

Further developments on these lines can be expected and the possibility of superimposing sign language encoded data on a basic human form, may be the direct analogy to the VIDAS proposal. Future examination of these issues will also need to have a clearer definition of the mobile terminal and the expanded screen necessary. It was not possible in this project to deal with this.

Chapter 8: Summary & Conclusions

Currently Available Video Technology

There are a large number of video communications products available today which have proved useful, using ISDN lines, the Internet, LANs and standard telephone lines. However, there are no suitable commercially available products based on the mobile telephone network. There are so many different factors involved (including resolution, frame-rate, colour depth, band-width, processing and storage requirements, delay etc.) that each product optimises some at the expense of others. The result is that the usefulness of each product is limited to a specialised application area, such as WWW 'thumbnails', or email videograms. Of the systems reviewed in this project, all the *software-based* encoding/decoding schemes produced unsatisfactory video, due to low frame rate, delay, or lack of clarity. This was the case both with Internet- and telephone-based products. The quality of video was not adequate to support comfortable, naturally signed conversation between deaf people. The *hardware-based* solutions are much more effective but also more expensive. One such system, PictureTel, was used in trials with deaf users to obtain quantitative results on the effectiveness of video communications for remote signing.

The video-conferencing trial using PictureTel showed that it is possible to sign at almost normal speed on a direct ISDN connection (2_64Kbs), although a higher degree of concentration is required for new users. However, remote finger-spelling is more difficult than face to face because of the limited screen resolution. Letters that proved hard to distinguish were L, M, N and R (all formed by placing fingers of the right hand on the left palm).

The Future

With the popularity of the World Wide Web and multimedia applications, much research is being carried out into online video: how the quality can be improved whilst at the same time reducing the software and hardware requirements. In general these gains will be made at the expense of algorithmic complexity – for example, some of the newer codecs make use of fractal compression (based on chaos theory) or neural networks. MPEG4 will be the new video standard, capable of handling high quality video at high bandwidths, as well as more acceptable video at lower bandwidths. The H263L video standard specifically targets low bit-rate bandwidths and should allow 28Kbs video of a quality which is currently only possible at higher bandwidths. Both MPEG4 and H263L will be finalised in November 1998.

Although mobile phones today are limited by a 9.6Kbs connection, *multi-slotting* will allow 28Kbs later this year. This research indicates that 28Kbs is the minimum bandwidth for acceptable video, so for the first time video communication based on mobile phones could be possible. In addition, microprocessors are becoming more powerful and more complex. The aim is to make the newer, more sophisticated algorithms viable in real-time and at low power, suitable for use in a battery-powered system.

The combination of state-of-the-art software and hardware will make mobile video conferencing a theoretical reality, but in practice a number of issues relating to the physical design will still need to be addressed: the size of the unit, the screen resolution, and the embedded camera. In view of the costs it will be some time before any particular product becomes well established.

Appendix 1

References

CCITT H.263 Image Compression (<http://www.ee.ethz.ch/~rmprince/h263.html>) - This site provides a summary of new features available in the H.263 standard, and includes MPEG encoded samples of the video sequences reproduced after H.263 encoding/decoding at various bit rates.

Robust H.263 Coding for Mobile Channels (<http://rice.ecs.soton.ac.uk/peter/robust-h263/robust.html>)

MPEG Home Page - <http://drogo.selt.stet.it/MPEG>

ACTS Information - <http://www.at.infowin.org/ACTS>

Details of current ACTS projects - <http://www.infowin.org/ACTS/ANALYSIS/PROJECTS>

MPEG general information - <http://www.MPEG.org/>

SNHC information - <http://www.es.com/MPEG4-snhc>

MSDL information - <http://www-elec.enst.fr/msdl/msdl.html>

ACTS central European Newsletter Special Issue on MPEG4 -

<http://www.esat.kuleuven.ac.be/~konijn/accents.html>

Compression Standards

ITU-T H.261

ITU-T H.263

ITU-T H.324

Currently accepted standard describing the operation of a terminal for video telephony.

The specified video telephone terminal operates using the following standards

ITU-T H.324/M

Extension of the H.324 standard, to achieve reliable operation on mobile networks.

ITU-T H.263+

This is the proposed short term enhancement to the H.263 standard, which is expected to be broadly similar to the H.263 standard, but to add a few new improvements and lead to a small increase in performance.

ITU-T H.263L

A further refinement to the H.263 standard, which was planned to be a longer term development, using new coding methods to achieve significant improvements over the existing H.263 specification.. Relevant documents relating to the progress of this standard are given in this Appendix .

H.324

H.324/M

MPEG1

MPEG2

MPEG4

Relevant ACTS Projects

(further information these projects can be found at <http://www.infowin.org/ACTS/ANALYSIS/PROJECTS>)

Emphasis (AC105) Architectures, Software and Hardware for MPEG4 Systems. The Emphasis project plays an active role in the definition of the MPEG4 standard, and is engaged in producing a software implementation of MPEG tools and algorithms, defining syntax and multiplexing of MPEG4 compressed objects, and producing specifications for enhancing processor architecture to meet the demands of MPEG4 applications. A platform independent software MPEG4 decoder is in development, although the compilation of the code is highly optimised for use on particular processors. A summary is reproduced in this appendix

FIRST (AC005) Flexible Integrate Radio Systems Technology The fundamental objective of this project is to demonstrate that it is feasible and cost effective to develop and deploy Intelligent Multimode Terminals capable of operation with multiple standards and with the ability to deliver multi-media services to mobile users. Employs work previously undertaken by the RACE MAVT project and the ACTS MOMUSYS project

MEMO(AC054) Multimedia Environment for Mobiles. Developing a generic architecture for the provision of interactive multimedia services to mobile and portable terminals. This is not particularly relevant to the type of applications we are interested in, as the main focus is on asymmetric systems where data is broadcast using the Digital Audio Broadcasting System with data rates of up to 1.8Mbit/s, with the low bit rate GSM link used only for interaction

MOMUSYS(AC098) Mobile Multimedia Systems Development and validation of the technical elements necessary to provide new audio-visual functionalities for mobile multimedia systems. Working in collaboration with international standardisation efforts, in particular MPEG4, developing techniques including robust tools and algorithms for content based video coding, and investigating techniques dealing with Synthetic Natural Hybrid Coding. Concentrating on scalability from very low bit rates to high bit rates. A summary is reproduced in this appendix

On The Move (AC034) Application Support Services for Distributed Mobile Multimedia

SCALAR (AC077) Scalable Architectures with Hardware Extensions for Low Bit rate Variable Bandwidth Realtime Videocommunications - investigating, developing and implementing a family of second generation coding techniques, with a focus on videotelephony applications. A summary is reproduced in this appendix

UMPTIDUMPTI (AC027) Using Mobile Personal Telecommunications Innovation for the Disabled in UMTS Pervasive Integration

VIDAS (AC057) Video Assisted With Audio Coding and Representation. Working on the development of a hybrid coding scheme to improve the quality of low bit rate coded video in videotelephone applications, by utilising the audio signal to extract information about the lip movements of the person depicted in the video, and using SNHC compliant methods to improve the quality of the video relating to lip movements. This project is being undertaken with co-operation from the Irish National Association for the Deaf, and is viewed as improving the usability of videotelephone systems for people with hearing impairments, possibly allowing lip-reading. The applicability of this method to our requirements is limited as phoneme extraction will obviously require a speaking subject. This is unlikely to be particularly effective in communication between deaf people, but may be useful for other situations.

Relevant documents

Coding of video conference stereo image sequences using 3D models(sigProcImComm 9(1997)125-135)

RACE R2072 MAVT (Race Mobile Summit 1995)

Advanced Video Coding Scheme For Low Bit Rate Communications (ACTS mobile summit 96 pp857 - 863)

ISO/IEC JTC1/SC29/WG11 N Short MPEG4 Description

ISO/IEC JTC1/SC29/WG11 N1251 MPEG-4 Implementation complexity considerations

ISO/IEC JTC1/SC29/WG11 N1375 MPEG-4 Verification Model & Core Experiments

ISO/IEC JTC1/SC29/WG11 N1495 MPEG-4 Requirements

Very Low Bit rate Audio-Visual Applications (Periera, Koenen)

Low bit-rate video coding using wavelet vector quantisation (IEE proc vis im sig proc 95)

Geometric Transforms (IEE proc vis im sig proc june96)

Very low bit rate segmentation based video coding using contour and texture prediction
(IEE proc vis im sig proc Oct95)
ITU Recommendation H.263
The MoMuSys Multimedia Terminal (ACTS 96 p850 - 856)
ITU Document LBC-96-136
ITU Document LBC-96-137
ITU Document LBC-97-029
ITU Document (multi-channel protocol)
ITU Document (error correction)
RACE Project 2053 MORPHECO
ACTS AC098 MoMuSys
ACTS AC077 SCALAR
ACTS AC027 UMPTIDUMPTI
ACTS AC057 VIDAS
ACTS AC105 EMPHASIS