Quality of Experience in Digital Mobile Multimedia Services

Hendrik Ole Knoche

A dissertation submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy of
University College London

10th August 2010

Declaration

I, Hendrik Ole Knoche confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Lausanne, 10th August 2010

Abstract

People like to consume multimedia content on mobile devices. Mobile networks can deliver mobile TV services but they require large infrastructural investments and their operators need to make trade-offs to design worthwhile experiences. The approximation of how users experience networked services has shifted from the inadequate packet level Quality of Service (QoS) to the user perceived Quality of Experience (QoE) that includes content, user context and their expectations. However, QoE is lacking concrete operationalizations for the visual experience of content on small, sub-TV resolution screens displaying transcoded TV content at low bitrates.

The contribution of my thesis includes both substantive and methodological results on which factors contribute to the QoE in mobile multimedia services and how. I utilised a mix of methods in both lab and field settings to assess the visual experience of multimedia content on mobile devices. This included qualitative elicitation techniques such as 14 focus groups and 75 hours of debrief interviews in six experimental studies. 343 participants watched 140 hours of realistic TV content and provided feedback through quantitative measures such as acceptability, preferences and eye-tracking.

My substantive findings on the effects of size, resolution, text quality and shot types can improve multimedia models. My substantive findings show that people want to watch mobile TV at a relative size (at least 4cm of screen height) similar to living room TV setups. In order to achieve these sizes at 35cm viewing distance users require at least QCIF resolution and are willing to scale it to a much lower angular resolution (12ppd) then what video quality research has found to be the best visual quality (35ppd). My methodological findings suggest that future multimedia QoE research should use a mixed methods approach including qualitative feedback and viewing ratios akin to living room setups to meet QoE's ambitious scope.

.

Acknowledgments

A large number of people accompanied me through the time of working on and writing this thesis that I feel indebted to. First and most Angela Sasse who as my supervisor has given me all the support, help and freedom one could wish for. I learned a lot more than about HCI and am grateful for having been part of her prosperous group. My second supervisor Mark Handley for valuable input during my research. My co-authors first and foremost John McCarthy who also acted as my mentor. Iain Richardson and Anthony Steed for the constructive feedback they gave me during my PhD viva and in the physical copies they received. Satu Jumisko-Pykköö for helping me prepare for my viva with some thought provoking questions, and for all the inspiring discussions at the various conferences we met. Albeit we could not find the time to write a paper together. Jens Riegelsberger, Will Seager, Sven Laqua, Hina Keval, Philip Bonhard, Philip Inglesant for being such inspiring colleagues from whom I learned various things. Nicolas Chuberre and Christophe Selier for helping leading the focus groups at Alcatel Space. Harald Weinreich for valuable discussions about early drafts of my thesis. Dimitrios Miras for his insights into video coding. Rob Kooij from TNO along with Tim de Koning, Pim Veldhoven who offered a different perspective and took the time and patience to evaluate my videos with VQM. Marco Papaleo for his skillful work on getting eye-tracking information of multiple participants overlayed onto a video - a functionality still missing from popular eye-tracking software. Marco's supervisor Giovanni Corazza for sending him to provide video material for the zoom study. Pablo Cesar and Dick Bulterman for a most fruitful and enjoyable collaboration in the later stages of my PhD. Shelley Buchinger for her tremendous energy and enthusiasm in exploring future funding possibilities and provision of baby clothing, which saved me a lot of time that I could in turn use for writing instead. Thanks to many anonymous reviewers that provided helpful feedback to sharpen my arguments and improve my scientific writing. Last but not least I owe Sarah Holsen a big thank you for supporting me throughout the process, transcribing a number of focus groups and continuing to help me improve my written English – a most wonderful time despite all the work. Nika Yola won the race and arrived before this thesis was finished. This thesis is dedicated to my parents, brother, wife and daughter.

Table of Contents

List of tables

List of figures

Glossary

AQ	Audio quality
AR	Angular Resolution in ppd or cpd
AS	Angular Size in degrees
cpd	cycles per degree, one black and one adjacent white line make for one cycle
CIF	common intermediate format, a video format with a resolution of 352x288
CSF	contrast sensitivity function
CVQE	Continuous Video Quality Evaluation, an objective video quality measure
D	viewing distance from picture or screen
dvd	design viewing distance
DVD	digital versatile disk
DVQ	Digital Video Quality metric, an objective video quality measure
DSCQS	Double-stimulus continuous quality scale
EPG	Electronic program guide
H	the height of a picture or screen
	a viewing ratio in which viewing distance is denoted in picture heights
HCI	human computer interaction
HDTV	High-definition TV
IPTV	Internet protocol TV, TV content provided through the Internet protocol
IS	information systems
ITU	International Telecommunication Union (-T telecommunication, -R research)
LS	long shot
MCU	medium close-up
MQ	Multimedia quality
MS	medium shot
MTF	modulation transfer function
p2p	peer to peer
ppd	pixels per degree
ppi	pixels per inch
PSNR	Peak Signal Noise Ratio
PVD	preferred viewing distance
PVR	preferred viewing ratio
	personal video recorder
PVS	preferred viewing size
QCIF	quarter CIF; a video format with a resolution of 176x144
QoE	Quality of Experience
QoS	Quality of Service
QVGA	quarter VGA resolution; 320x240
SDTV	Standard Definition TV

SQRI	Square-root integral
SSCQE	Single Stimulus Continuous Quality Evaluation
SSI	structural similarity index, an objective video quality measure
TV	television
TPM	task performance measure
UCD	user-centred design
UX	user experience
VA	visual angle
VD	viewing distance
VGA	Video graphics adapter standard synonymous with 640x480 resolution
VLS	very long shot
VQ	Video quality
VQM	Video Quality Metric, an objective video quality measure introduced by Pinson & Wolf
VR	viewing ratio, the viewing distance divided by H
VX	Visual experience
WMA	Windows Media Audio format
WMV	Windows Media Video format
XLS	extreme long shot

Chapter 1

Technology is a way of organizing the universe so
that man doesn't have to experience it.
– Max Frisch

Introduction

People enjoy multimedia content and can choose from a range of ways to consume it on portable devices: physical media (DVDs), downloaded and stored content or receiving it wirelessly from broadcast or mobile telecommunication providers. Mobile TV, i.e., consuming audio-visual content on mobile devices, was touted the emerging killer application of the 21st century (Kumar 2007). Broadcasters worldwide have run trials for mobile TV based on adaptations of the digital TV broadcast standards and a number of them have launched broadcast services. In competition with those, mobile network operators offer digital content through 3G and higher networks. By 2012, the worldwide capital expenditure of mobile operators is predicted to exceed $150 billion, to improve coverage and to rollout advanced services e.g. mobile multimedia application like mobile TV. In countries with rolled out mobile TV services, uptake lags behind expectations. One possible explanation for this blames consumers' reticence to invest in specialized devices in general (Meyer 2007) and competing standards in particular (Briel 2008). Another possibility is that of a more fundamental problem: that customers are sceptical whether the delivered experience on small mobile devices is worthwhile in the first place (Deloitte 2006). Apart from the restrictions in terms of reception, battery life, programme lengths and the delay in the delivery of live content, the user experience of mobile TV is dominated by the visual experience of the content, which is affected by the limited presentation space and the lower resolution compared to regular TV. High content production costs and customers' reluctance to pay a premium for tailor-made mobile content (KPMG 2006) mean that most mobile TV content consists of transcoded broadcast TV content. With first mobile TV rollouts being cancelled due to insufficient uptake service providers are looking into offering services that are both financially viable while providing an adequate Quality of Experience (QoE) to their customers. Research on QoE is still in its infancy. It currently draws heavily on video quality assessment methods whose applicability to predict user preferences and maximised QoE is questionable (the discussion of this can be found in Sec. 2.6.7 and 2.6.8). This thesis describes my research on the key factors that determine QoE in mobile multimedia applications, with a focus on mobile TV.

The substantive results of my research can help service providers decide how much content to deliver, under which conditions, at what qualities, and which enhancements should be made to content that was originally produced for regular TV. My research covered low resolution and low encoding bitrates a parametrical sub-space that service designers face in a competitive mobile entertainment market.

On the methodological side, my results show that a binary assessment method – *acceptability* - provides better prediction of peoples' preferences than objective analysis of video quality. The highest levels of acceptability coincided with users' preferred conditions when trading of resolution for size.

1.1 Research problem

Mobile multimedia applications allow users to consume content, but there are limitations in terms of where, when, for how long and how they can interact with and experience the content. The applications present different kinds of content through a digital device that recreates sensory input in the auditory, visual and textual domain. Now that mobile multimedia applications and services like mobile TV have become technically feasible on mobile phones, portable game platforms, music players, laptops and other mobile devices, service providers focus their attention on how to best deliver these services. Many constraints are imposed by the diversity of devices on which people consume or interact with multimedia content e.g. size and resolution, and a range of trade-offs have to be made faced with the bandwidth limitations imposed by wireless content delivery. Mobile multimedia service deployment requires large up-front investments from service providers who currently lack an understanding of what constitutes a good QoE of these novel services on such a diverse array of target devices. To plan services and their pricing, service providers need to understand consumers' acceptance, uptake and continued use of the service and how this depends on the service's QoE. This problem is exacerbated by the fact that many service providers have little knowledge about multimedia content production, while content producers typically only target a single device - standard definition TV.

At the beginning of this thesis both substantive and methodological knowledge was missing. In the field of mobile multimedia in general and mobile TV in particular it was unclear:

1. what constitutes subjective QoE for mobile multimedia services, how do its constituting factors relate to the service providers design decisions and
2. what are meaningful and reliable measurements of QoE?

Almost no research existed either on sub-TV resolution content or on mobile devices for passive consumption. Research and video quality models were based on standard and high-definition TV content, but used a) unrepresentative viewing distances, and b) often not even moving images (still pictures were used). Previous multimedia research focused on identifying configurations resulting in maximum objective or subjective video quality, assuming that these regardless of e.g. size and immersion would coincide with user preferences. Ratings of an experience need to correlate with and reflect people's preferences.

The visual experience needs to be measured in a way that can be used to predict QoE. Most video quality studies collect quantitative ratings only and leave valuable information unearthed as to what assessors' ratings are based on. The research problem consisted of identifying the factors that shape people's experience in consuming audio-visual content on mobile devices and reliably quantifying their contribution through appropriate methods under ecologically valid conditions.

1.2 Research scope

This research approaches the field of television research from a computer science angle. Television engineering and standards have a long research tradition: to use the limited frequency spectrum efficiently, technical solutions and standards were based on the investigation of the human visual system and its limitations. Television sets are hardwired in terms of what signals they can display. Together with the economies of scale in production this has resulted in a small set of standards that have been used for decades. Content producers have specialized in content production within the constraint set imposed by the standards. Computer science has taken a different approach: video coder and decoder combinations (codecs) can be set to render content at different resolutions and encoding bandwidths. Multimedia applications and services such as Joost and YouTube employ the codecs to constrain the media encoding to match the technical infrastructure. In many cases the technical trade-offs - especially in relation to the content, which was not produced with these constraints in mind - are not informed by users' preferences, needs and value perception. A lack of user-centred design is apparent in the design decisions of appliances, software applications (Cooper *et al.* 2007) and internet provided services (Bouch *et al.* 2000). User-centred design aims to understand user needs in given contexts, and to produce design solutions that meet those needs and the constraints imposed by the available technologies. Within a typical context of use - for example while travelling on a busy over ground train - many factors determine the experience of an application including the interaction through the user interface (UI). The research scope of this thesis is limited to a subset of QoE, namely the factors that influence the users' visual experience of the service while it is being used on a mobile device, and therein concentrating on the threshold at which the experience becomes acceptable.

Most of the research presented here focuses on the experience of TV content delivered through wireless networks to mobile devices. Compared to the wealth of research on human perception of video, its quality and its storage, retrieval and networked delivery over fixed networks, there has been little research on mobile multimedia services and its QoE. However, the majority of work in multimedia research has been carried out in bandwidth-limited environments. This body of work informed the research on which this thesis is based and the literature review, therefore, is grounded in the many studies that have investigated multimedia delivery through packet networks. In mobile environments the frequency spectrum and therefore bandwidth is scarce but the number of users and the amount of content, i.e. data they are interested in, is growing steadily. Therefore, resources will be limited, differentiated through pricing - or both, for the foreseeable future. This will limit the audio-visual quality of mobile TV content. Furthermore, due to the high cost of content production, mobile multimedia services such as mobile TV will continue to use footage that was not originally produced for low resolution devices. Methods for subjective video quality assessments have been standardised by the International Telecommunications Union (ITU), but I will provide evidence in the background chapters (2-4) that applying these standards is not sufficient to predict QoE. This research reviews the existing methods for video quality assessment, and justifies the use of a recently adopted method to find out which video quality profiles are acceptable to users in a specific mobile context.

My main research question was:

1. Which factors affect the QoE of mobile multimedia services, specifically mobile TV, and how exactly?

The supplementary research questions are:

2. Which of the existing methods are suitable for establishing and measuring these factors?
3. Under which conditions does content ported from TV to mobile devices result in a satisfactory visual experience?

This includes the following detailed research goals (RG):

1. Identify encoding bitrates of diminishing return of different content types.
2. Understand the contribution of size and required sizes in relation to viewing distance.
3. Identify required resolution.
4. Identify possible problems with shot types, audio and text.
5. Validate current objective and subjective models.

This research did not investigate all effects pertaining to the capture (such as colour aberration, saturation and noise) and delivery of moving images through digital networks (e.g., packet loss and bit errors). All of these influence the visual experience and have to be considered in the overall system design trade-off. However, an investigation of the contribution of loss and errors should be grounded in knowledge of how the fundamental parameters size, resolution and shot types - addressed in research goals 2-4 - affect the baseline of visual experience. By itself, my focus provides insights for the visual experience of content delivered free of errors common for physical media and many on-demand services.

1.3 Research approach

This research approaches the problem from a user-centred design (UCD) point of view. In UCD, designers of services consider the users' interactions with a service at various stages of the development cycle, and evaluate and validate their design decisions of a service with real users. It considers users in their typical physical and social context when interacting with a service to achieve tasks or more general goals. User needs and expectations guide the design of a service in this approach. The challenge is that prospective users are not very good at predicting if and how they might use new services. Without having seen or experienced a service first hand their opinions represent educated guesses. Envisionment techniques (Ehn & Kyng 1991) such as mock-ups, prototypes, demonstrators and other artefacts with varying degrees of resemblance to the actual service are used at the various stages of the service inception to enable users to picture possible interactions between them and the new services. Furthermore, different user groups may be interested in different services and content. In order to obtain ecologically valid results, all the presented studies used participants who had interest in specific services and content they were asked to judge.

At the start of this research in 2004, service providers in Korea had just deployed Satellite Digital Media Broadcast (S-DMB) services that allowed for mobile TV watching on mobile phones, but there was no published research on user responses to them. Only a couple services that could be described as a form of mobile TV existed in the UK, but they offered multimedia content to download from gallery lists by means of point-to-point connections through 3G networks. One previous large-scale study evaluated mobile TV use in Finland: participants had used tablet PCs and PDAs to experience mobile TV that was broadcast like regular TV with different channels (Södergård 2003). However, the prevalence of the mobile phone as a platform required an evaluation of people's expectations and attitudes towards mobile TV service delivered onto their mobile phones.

The initial step was to find out more about users' current experiences with their mobile phones and services, as well as expectations and needs about future services, especially mobile TV. To start with, I conducted a series of focus groups in two countries, in which participants learned about possible content types and services they might be able to use. From the qualitative analysis of these results and a thorough background research of the existing literature, the gaps in the existing body of knowledge were identified, and research to fill them planned. Most surprisingly, the contribution of size to visual experience had only been a subject of research for immersive displays (see Sec. 2.3.8) or had been confounded (see Sec.2.6.7) with resolution (RG2). Both subjective and objective video quality models focussed on video quality per area irrespective of size (RG5) and they typically were based on viewing ratios uncommon to the living room - their real use context (see Sec 2.6.6). Although a major factor in video quality, the role that resolution of both the screen and the content plays on visual experience (see Sec. 2.6.8) was inconclusive (RG3). Video quality studies also often ignored the influence of audio and text (RG4). Finally, appropriate encoding bitrates for the display of TV content on small screens at low resolution were unknown (RG1).

To address my empirical research goals, I conducted a series of lab studies investigating parameter combinations typical of mobile devices in terms of size, resolution, encoding bitrates in both audio and video for content types that appealed to people in mobile use contexts as previously identified by focus groups. Each study was an independent investigation of one or several factors of QoE, and hypotheses, methodological details, results, and a discussion of the findings are presented for each of the studies. The motivation and focus of subsequent studies was usually informed by the qualitative and quantitative results of the preceding studies as indicated by the white arrows in . Many parameters such as encoding bitrates, sizes, resolution, content types and actual video clips were re-used in several studies along with the method to allow for comparison between studies. All but one study employed the measure of acceptability (McCarthy *et al.* 2004a) – an approach for the assessment of QoE, which allowed for comparisons between studies, increased the validity of the results and helped to test the method under different circumstances. The experiments included semi-structured debriefing interviews at the end of each session to gain an understanding of how the factors contributed to the results, or if there were any interactions between them. The analysis of the qualitative feedback then informed subsequent studies (*cf.* Sec. 1.4). All but study 5 presented content on real mobile devices. Furthermore, I compared my lab results with in-situ results from a field study, in which users watched content while travelling on public transport. The results of the first study were cross-compared results with the objective video quality measure VQM (provided by the Dutch research body *TNO* in Delft). The results from the shot-type study were compared with results obtained through an objective quality measurement *peak signal-to-noise ratio* (PSNR). In summary, I used a multiple method approach and built upon the findings from preliminary studies in follow-up studies to further disambiguate the results and arrive at a better understanding of the QoE of mobile multimedia services.

1.4 Overview of the thesis

This thesis focuses on how to best present audio-visual material for consumption on mobile devices. depicts the structure of this document. Chapter 2 presents a critical review of the literature on the recreation of audio-visual content based on the limits of human visual perception. It focuses on the design trade-offs for capture, storage, delivery and presentation of video and how many innovations in entertainment technology were guided by improving its visual experience. Chapter 3 reviews the methods that have been used to measure user experience – at the intersection of human-computer interaction and motion picture engineering. It details several limitations of existing methods to assess video quality and motivates the use of a mix of methods that I used in my studies. Chapter 4 takes a closer look at the *status quo* in mobile television research and the methods that were used.

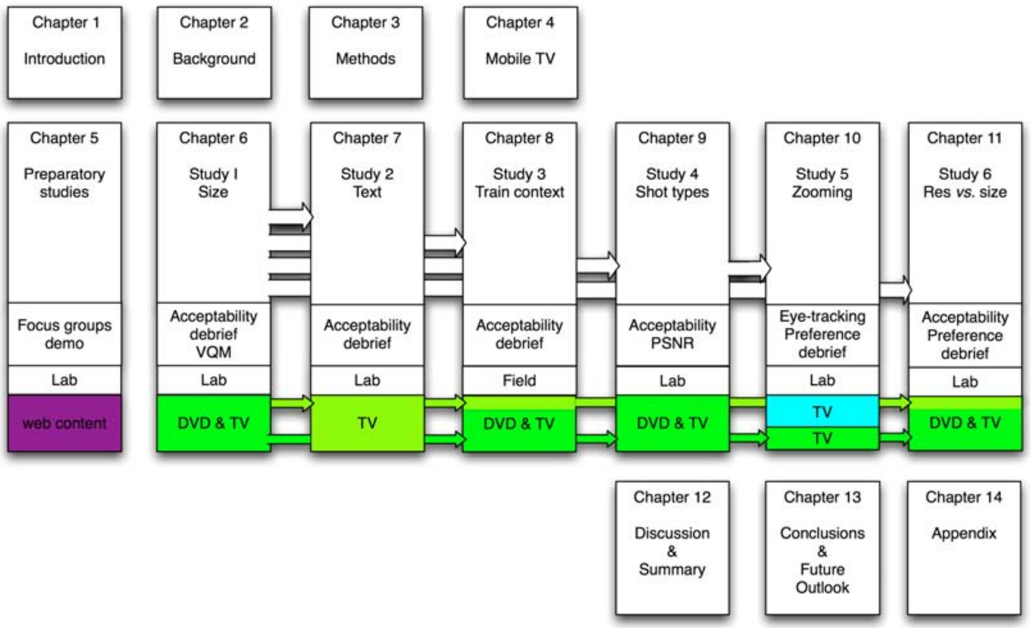


Figure 1: Thesis structure

The depictions of chapters 5-11 include the study name, focus, used methods, study type and the used content. White arrows indicate how studies informed subsequent ones, coloured arrows indicate the re-use of material

Chapter 5 sums up the findings of a series of focus groups I conducted in addition to the previous mobile TV research to guide the design and choice of methods in Study 1. Chapter 6 presents Study 1, which assessed the effects of size, encoding bitrates of different content types and the effect of audio on the acceptability of video quality at a nominally constant angular resolution. The qualitative results of the first study revealed that the legibility of text was key factor in participants' assessment of the acceptability of the video quality. Participants also complained about missing details in the very long shots of football content. Consequently, two follow-up studies specifically addressed the effects of text quality (study 2) and shot types (study 4) on acceptable video quality settings. Chapter 7 presents study 2, which looked at the contribution of the visual quality of text to QoE. The study was prompted by numerous complaints about size and legibility in Study 1. Chapter 8 describes Study 3, which replicated the design of and reused content from Study 1 and 2 – both conducted in a controlled setting – on the London underground and therefore assessed the ecological validity of the results obtained under lab conditions. Study 4 presented in Chapter 9 re-analyses the results of Study 1 - based on the shot types used - and compares this to results of the popular objective video quality measure peak signal-to-noise ratio (PSNR). Chapter 10 describes Study 5 on people's preferences and limits of zooming into shot types depicting people from afar. Study 4 had identified these extreme long shots as problematic. Study 5 used a combination of eye-tracking, subjective preference measures and individual interviews to arrive at the presented results. Chapter 11 describes the concluding study 6 that addressed the trade-off between size and angular resolution of video. This study disambiguates the results from study 1 in which size and resolution were confounded. Chapter 12 holds the discussion of the thesis, which pulls together the results from all studies and suggests a future research agenda in the field of audio-visual content consumption on mobile devices. It summarizes the substantive empirical, theoretical, and methodological contributions and provides recommendations for researchers and practitioners. Chapter 13 presents the thesis conclusions in relation to the research goals.

1.5 Thesis contributions

The contribution of my research includes a comprehensive synthesis of previous research in the field of multimedia services and an extensive set of both substantive and methodological findings. It synthesizes the existing knowledge about TV and picture quality in relation to various parameters. It also presents the trade-offs people make when maximizing their visual experience under lab conditions and in the field. The results bridge the gap between what users experience, prefer and value when consuming video content on mobile devices and the quantitative technical parameters that describe a service.

The substantive contribution lies in the identification and systematic evaluation of factors affecting the perceived acceptability of visual experience in mobile multimedia applications. These factors include size, resolution, content types, encoding bitrates, audio quality, text quality, shot types, and zooms. People expect a similar relative picture size (expressed by the viewing ratio - the viewing distance divided by the screen height) for mobile TV to what they are used to in typical living room setups. At typical viewing distances this required a minimum picture height of 4cm in an indoor lab setting and 4.5cm in a field setting on a train. The resolution of the content can be greatly reduced and content encoded from QCIF resolution (176x144 pixels) onwards can provide an acceptable experience at adequate sizes. Displays with high resolution affect the visual experience positively and the content can be up-scaled on them to the point at which the angular resolution of the content in the eye of the beholder is reduced to 12 pixels per degree (ppd). Along with previous research on viewing preference of TV content my findings suggest that an angular resolution between 14ppd and 11ppd represents a general threshold below which the acceptability of the visual experience of video rapidly drops off. However, for the best visual experience viewers of QCIF content prefer a higher angular resolution of around 19ppd. This angular resolution is also sufficient for the rendition of text (of 19 arc minutes height) included in TV content given sufficient encoding bitrates. When delivering TV content at QCIF resolution to mobile devices of at least 4.5cm screen height content adapatation in terms of the used shot types is not required apart from extreme long shots in football content. Players in this content should be at least 0.8° in height to be acceptable to football fans. The end user can achieve this by zooming (manually or software assisted) or by using devices with large enough screens; the content producer can help provide this acceptable height by creating zoomed-in content. The contribution of picture size to the visual experience is much larger than its contribution to video quality. Whereas an acceptable visual experience of mobile TV starts with 4cm picture height and 14ppd angular resolution, size only starts affecting the perception of picture quality once an angular resolution of 35ppd is reached. My findings refine the concept of QoE in the domain of mobile multimedia applications by promoting the concept of visual experience. My results point to a set of improvements that can be made to mobile multimedia services.

The methodological contributions of the thesis lie in the successful application of the concept of *acceptability,* through the further development of QoE assessment across a number of studies in a novel context as part of a mixed methods approach. The method for determining acceptability

a) required little effort from users in expressing satisfaction thresholds of their visual experience,
b) provided more meaningful results than existing methods,
c) was successfully applied both in lab and field settings and
d) was compared to other subjective and objective measures.

In my mixed methods approach I coupled acceptability and other measures like preferences with qualitative feedback. Feedback gathered directly while participants were watching and assessing the acceptability of video clips helped to disambiguate the contribution of confounded factors to the visual experience – in my case, size and resolution. Future research into QoE should employ qualitative methods to discover further parameters that affect QoE.

1.6 Research context

The research presented in this thesis was performed as part of two European Commission-funded projects. The first - MAESTRO - aimed to provide mobile TV through a satellite infrastructure called SDMB, which used carousel re-transmission and terrestrial gap-fillers to enhance coverage in urban areas. The service could provide live channels and caching of various programmes on non-volatile storage media. Pre-cached content could be consumed at any time without satellite reception. The second project - UNIC - provided broadband interactive services through set-top-boxes on TV screens. In this approach, mobile devices acted as *secondary screens* to the TV set-top-box infrastructure. The set-top-box would receive content through broad- or multi-cast protocols and relay them to the mobile device for independent viewing.

1.7 Publications

The research presented in this thesis has lead to a number of publications the most important of which are listed in Table 1 and mapped to the contributions of the thesis. As primary author of all of them, I have included my contributions to these publications into my thesis and generally excluded my co-authors' contributions or in a few cases have explicitly attributed this to them.

Table 1: Selected publications relating to this thesis

	Contribution	*Publication*
Chapter 4	Overview of mobile TV requirements	Knoche, H., Sasse, M. A. (2008) **Getting the big picture on small screens: Quality of Experience in mobile TV.** In Ahmad, A. M .A. & Ibrahim, I.K. (eds.) *Multimedia Transcoding in Mobile and Wireless Networks*, pp. 31-46, Information Science Reference
Chapter 5	Initial mobile TV user requirements	Knoche, H., McCarthy, J. D. (2005) **Design Requirements for Mobile TV.** In *Proceedings of Mobile HCI 2005* , pp 69-76
Chapter 6	The effects of bitrate, audio quality and size on acceptability	Knoche, H., McCarthy, J. D., Sasse, M. A. (2005) **Can Small Be Beautiful? Assessing Image Size Requirements for Mobile TV**. In *Proc. of ACM Multimedia 2005*, pp. 829-838
Chapter 7	The effect of visual quality of text on acceptability	Knoche, H., McCarthy J., Sasse, M. A. (2006) **Reading the Fine Print: The Effect of Text Legibility on Perceived Video Quality in Mobile TV.** In *Proc. of ACM Multimedia 2006*, pp. 727-30
Chapter 9	The effect of size on different shot types	Knoche, H., McCarthy, J. D., Sasse, M. A. (2006) **A close-up on Mobile TV: The effect of low resolutions on shot types.** In Proceedings of EuroITV, 25-26 May, Athens, Greece Knoche, H., Sasse, M. A. (2008) **How low can you go? The effect of low resolutions on shot types.** In *Personalized and Mobile Digital TV Applications in Springer Multimedia Tools and Applications Series,* Vol 36(1-2) pp. 145-66
Chapter 10	Optimal zooms into extreme long shots	Knoche, H. Papaleo, M., Sasse, M. A., A. Vanelli-Coralli (2007) **The Kindest Cut: Enhancing the User Experience of Mobile TV through Adequate Zooming** in *Proc. of ACM Multimedia 2007*, pp. 87-96
Chapter 11	Preferred size and angular resolution and their limits.	Knoche, H., Sasse, M. A. (2008) **The sweet spot: How people trade off size and definition on mobile devices**, in *Proc. of ACM Multimedia 2008*, pp. 87-96 Knoche, H**.**, Sasse, M. A. (2009) **The Big Picture on Small Screens: Delivering acceptable video quality in mobile TV.** In *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMCCAP), 5(3): article number 20

Chapter 2

The cinema is an invention without a future
- Louis Lumière

Background

The requirements for pleasurable consumption of multimedia content lie at the heart of this thesis. The multiple media involved in my case include video, text and audio but my focus lies on the visual domain. The mediation of moving images includes the stages of capture, editing and delivery and display on mobile devices. In order to understand the requirements for moving images one needs to understand the thresholds and limits of human visual perception, the factors that affect the visual perception of moving images and how the design of video technology have been guided by the human visual system. The thesis will thus review how the different stages in mediation influence these factors, and how engineers made trade offs in the past. Due to the lack of scientific results on mobile video consumption, I reviewed research on standard -and high-definition broadcast TV, as well as computer-mediated communication, which operates in more resource- constrained settings. Although many of the parameters - such as viewing distance, viewing ratio, angular resolution and video resolution - are interrelated, many studies focussed on a single factor – this makes comparisons difficult, especially when descriptions of the other factors were not provided. I will therefore group previous results into separate subsections, depending on their main angle of inquiry. I include an overview on how video content can be adapted to a heterogeneous set of mobile devices and summarize the state of the art findings on the most relevant factors, which provided the basis for the factorial designs of my experimental studies.

People's perceptions, preferences and assessments lie at the heart of my inquiry. Research in psychology and human perception tends to label the people in its studies *subjects*, studies on discretionary computer use commonly prefer the term *participant* to include the notion of choice (Grudin 2005), people interacting with technology are referred to as *users* and in video quality centred studies *assessors* provide ratings of quality. I will use these terms throughout this thesis to indicate the angle of inquiry of a reported study.

2.1 Key terms

I will first briefly introduce the key terms addressed in this thesis and then explain them in more detail in their corresponding subsections. A scene (*cf.* Figure 2) is captured with a sensor of a given resolution (16x12pixels) and aspect ratio (4:3). Viewing the captured scene (content) on a display of height *H* from a distance *D* results in a viewing ratio *VR.* Similarly, the relation between D and H can be expressed through the angle θ that the display subtends in the eye of the beholder, which is referred to as the *angular size* or *visual angle* of the screen. Many displays have a resolution (*addressability*) different from the captured content and allow the user to display content non-natively - a pixel in the captured content is mapped to more (up-scaling) or fewer (down-scaling) pixels on the display. I use the term *preferred* *viewing size* (PVS) to refer to when users adjust the display of the content on the device in this manner to optimize their viewing experience. For example, in order to fill the whole screen of height *H* and display resolution of 32x24 pixels the captured content needs to be up-scaled by a factor of two in Figure 2. The resolution of the content remains 16x12pixels and the angular resolution (AR_c) of the content in pixels per degree (*ppd*) is 12pixels/θ. The display's angular resolution (AR_d) is 24pixels/θ. Throughout this thesis angular resolution (AR) refers to the angular resolution of the content unless stated explicitly otherwise. Viewers who want to experience a different picture size can also change the visual angle of the screen by adjusting the distance between their eyes and the screen to their *preferred viewing ratio* (PVR) or more precisely to their *preferred viewing distance* (PVD). With mobile devices this can easily be done by e.g. pulling the device closer or a change of seating posture whereas standard TV screens require the viewers to move their bodies and possibly furniture closer to the screen. While a reduction of *D* reduces the *VR* and increases the *angular size* it reduces the angular resolution of the content (and the display).

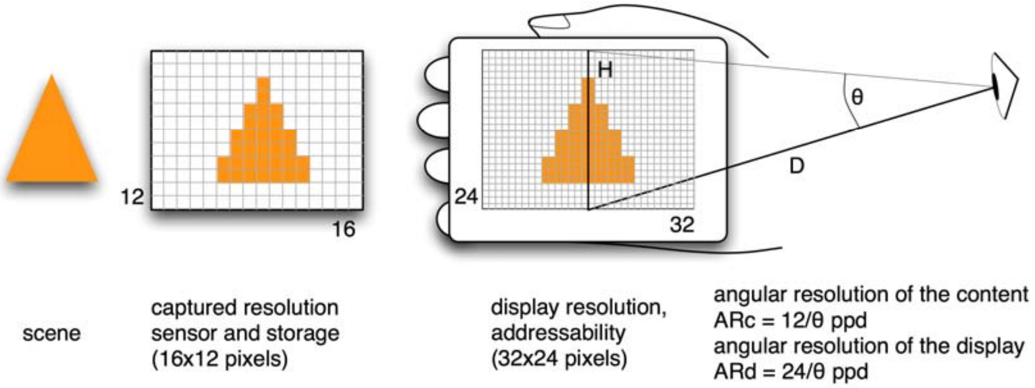

12
16
H
θ
D
24
32
scene
captured resolution sensor and storage (16x12 pixels)
display resolution, addressability (32x24 pixels)
angular resolution of the content
ARc = 12/θ ppd
angular resolution of the display
ARd = 24/θ ppd

Figure 2: Resolution of captured content, its display on a display with higher resolution (addressability) and the angular resolution of content and the display

2.2 Mediated visual information

The recreation of occurrences in the real world through visual and audio-visual media has fascinated and entertained people for a long time. The process of mediating visual information involves a number of stages, a simplified model of which is depicted in Figure 3. Light that reflects from objects within the view of the camera is captured by a lens and projected on a sensor medium, from hereon referred to as the sensor. The objects in the camera view might move and the camera may introduce further motion in the captured *scene* through pans and zooms. The sensor is a photosensitive surface, i.e., sensitive to light. The read out from the sensor allows scene images to be recorded successively to some form of storage or to generate a signal. For film cameras sensor and storage are nearly identical[1] but for digital recording systems the information gathered by the sensor is stored in a representation in non-volatile storage. The material can then be edited and delivered to a device that includes a display. On raster displays, the delivered image is then recreated by means of small picture elements called TV line pairs or pixels in the computer domain. At the end of this process the mediated version is perceived by the human visual

[1] The film requires developing.

system. Due to technical imperfections and limitations distortions may be introduced at any of these stages that were not present in the original scene and degrade the experience of the viewer.

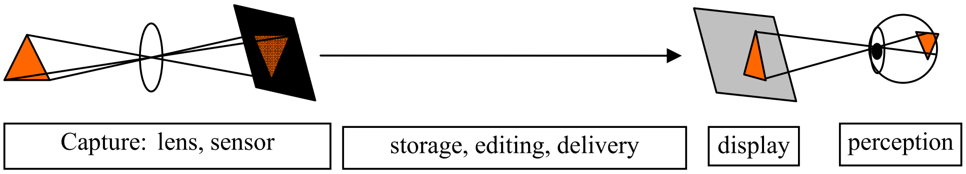


Figure 3: Mediated communication stages - from capture to perception

The first moving pictures were shown as attractions at fairs and later in movie theatres. Film cameras captured the footage on film. The film material was edited and copied in such a way that it could be shown to potentially large audiences by means of projectors on reflective screens. At first, presentations had no sound and lower frame rates to save on film stock while shooting (Poynton 2003). After the proliferation of television use in the home technology in cinemas evolved to widescreen colour presentation and high quality surround sound. For analogue TV this chain was standardised - the cameras that recorded a scene and delivered their representation in a signal resulted only in correct depictions on a TV set of the same standard. In digital systems the requirements for how content is captured, encoded, stored, processed, packetized for delivery over a network, decoded and viewed on a viewing device are more relaxed. But many current systems still partly employ equipment based on old analogue standards. However, production and delivery standards do not have to be identical (EBU 2004). Mediation of imagery has always factored in the limitations of the human visual system. The next sections present the capabilities and limitations of human visual perception of moving images.

2.3 Human visual perception

Human vision provides individuals with the ability to resolve visual detail in colour and visual changes of objects from a distance. As illustrated in Figure 4 light enters the eye, gets refracted by the cornea and lens and falls on the retina – light-sensitive tissue that lines the back of the eyeball. Photoreceptor cells called rods and cones that are located on the retina respond to light of wavelengths between 350 and 750nm (Schwartz 2004). The impressions obtained by the eye are transmitted to the visual cortex of the brain and result in visual percepts. The importance of vision is not only mirrored in the proportion of the human brain that is devoted to it. Vision can dominate other senses; for example, when conflicting information is provided one's experience can follow what the eyes have perceived, e.g. in the ventriloquist and McGurk effect (1976).

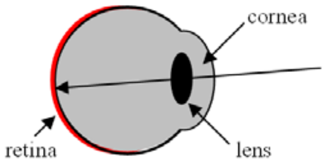


Figure 4: A simple model of the eye

2.3.1 Accommodation and vergence

The human visual system uses two mechanisms to target and focus on objects located at different distances from it: convergence and accommodation. Convergence refers to the inward movement of the eyes when looking at nearby objects while accommodation describes the focusing on objects of different distance by means of physically deforming the lens of the eye. The default distance at which objects appear sharp when opening the eyes is around 75cm for young people and farther away for old people.

This is called the resting point of accommodation (RPA). The resting point of vergence (RPV) is 115cm when looking straight (*cf.* Figure 5) and decreases to 90cm when looking 30 degrees down – a common posture when watching content on a mobile display (*cf.* Figure 6). Resting points change over time when continuously focusing on objects nearby.

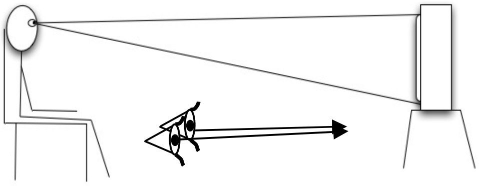


Figure 5: Watching TV on standard screen with straight viewing angle

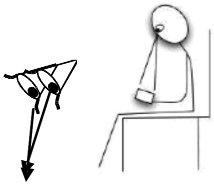


Figure 6: Watching on mobile device with a downward angle

2.3.2 Limits on viewing distance

Both accommodation and vergence might strain viewers' eyes (Collins *et al.* 1975), (Fisher 1977). The stress of convergence contributes more to visual discomfort than the stress of accommodation. No studies have shown greater fatigue with viewing distance further than the RPV but continued viewing at distances closer than the RPV can contribute to eyestrain according to Owens & Wolfe-Kelly (1987). When viewing distances approach 15cm, people experience discomfort (Ankrum 1996) due to convergence and accommodation.

Research on home viewing distances of standard definition TV (SDTV) has shown that spatial limitations and the layout of the average living room prompt people to watch TV at the so-called Lechner (US, 9ft) or Jackson (Europe, 3m) distance of approximately 9 feet or almost 3m (Diamant 1989). Unfortunately, both of these values are poorly documented - their original sources are not readily accessible. As recently as 2004 Tanton reported a median viewing distance for SDTV of BBC employees of 2.7m (8.5*H*). As Poynton (2003) argues, the viewing distance in the home could be considered fixed. As I will show later in Sec. 2.6.6 and Sec. 2.6.7 this is not considered in current recommendations for TV video quality assessment.

2.3.3 Visual angle and viewing ratio

The visual angle (VA) of an object of height *H* describes the angle *θ* the object subtends at the eye of the beholder. It is also referred to as *angular size* and depends on both the viewing distance *D* and the size of the object. Visual angles are expressed in degrees and minutes and seconds thereof. The visual angle can be calculated by the following equation:

$$\theta = 2\arctan\left(\frac{H}{2D}\right) \qquad \text{Eq. 1}$$

The astronomers' rule of thumb for 1° is the subtense of the nail of the small finger when held at arm's length (Poynton 2003). For practical purposes the viewing distance *D* is measured from the cornea of the eye but for optical calculations the distance from the visual centre of the eye - the first nodal point as depicted in Figure 7 - is preferred (Ware 2000). In television research the visual angle of the whole screen is commonly expressed in terms of the *viewing ratio* (*VR*) - the ratio between the viewing distance (*D*) and the image height (*H*) of the visible screen area:

$$VR=D/H. \qquad \text{Eq. 2}$$

Some researchers refer to the viewing distance in multiples of the picture height. A viewing distance of 5*H*, for example, means that the distance between the viewer and the screen is five times the height of the screen. Both VR and VA could be expressed in terms of width or the diagonal instead of the height. In fact, in early television consumer research (McVey 1970) it was common to describe viewing distances in terms of picture widths. Although the use of picture heights is more common nowadays the use of picture widths might be more appropriate since wide-screens are becoming more popular (see Sec. 2.6.1) and the human visual field is wider than it is high.

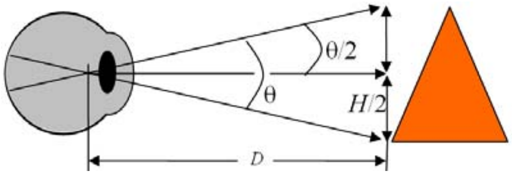


Figure 7: Viewing distance (*D*), object height (*H*) and the visual angle θ

2.3.4 Human visual acuity - spatial resolution

The amount of detail resolvable by the human eye is primarily limited by the density of the light-sensitive rods and cones on the retina. The highest acuity is in the *discriminatory visual field,* an area of roughly 3º from the visual centre of the eye. In the surrounding effective visual field the visual acuity is reduced to 10% but small figures can be discriminated quickly (Engle 1971). Ophthalmologists distinguish between three types of visual acuity: minimum visible acuity, minimum resolvable (ordinary) acuity, and minimum discriminable acuity (hyperacuity) (Westheimer 1992). Most frequently used within the engineering literature is minimum resolvable (ordinary) acuity. This is determined by a person's ability to identify a target – such as whether a letter is a C or an O – and depends on identifying the presence of a gap or feature in the letter. According to (Luther 1996), normal 20/20 vision is classified as the ability to resolve 1 minute of arc (1/60º) and translates to 60 pixels per degree (ppd). The numerator in 20/20 vision represents the distance to the target and the denominator the distance at which average people can identify the target. The maximum number of pixels p_{max} that can be resolved by a human at a given distance *d* and a picture height *h* can be computed by the following equation:

$$p_{max} = H/D \times 2 \tan(1/120) \qquad \text{Eq. 3}$$

Low luminance (lighting) and low contrast reduce human spatial resolution. The interaction between contrast and spatial resolution is described by the modulation transfer function (MTF) (Ware 2000), which I will address in more detail in relation to capturing in Sec. 2.4.1. Many acuity limits of visual resolution are measured through tasks and assume ideal conditions with sufficient luminance and contrast or signal to noise ratio. Scott (1955) defined the demand modulation function (DMF) for observers to detect targets made up of three bars.

In practice people do not require as much detail as their visual acuity would allow them to resolve. Reduced resolution is common in nature e.g. with mist or haze and softeners that remove high spatial frequency information are used for artistic reasons in pictures and video. According to Birkmaier (2000) moving images of approximately 22 cycles per degree (44ppd) are perceived as having a sharp image. This value is achieved when an SDTV display is viewed at 7*H*.

2.3.5 Requirements for the readability of text

Text appears frequently in TV content, especially in news programs - producers often use subtitles, headlines and ticker lines. Previous studies on text legibility in HCI research have examined the various dimensions, e.g. contrast & colour, formatting, size, and dynamism, all of which influence reading performance on computer screens (see Bergfeld Mills & Weldon (1987) for an overview: for a specific review on the effects of text size on legibility on computer displays see Smith (1979)). Larger text is not necessarily easier to read: research on printed text showed that the optimal reading performance is usually achieved with fonts of 9- or 10-points (Tinker 1963). Fonts larger and smaller than this resulted in a decrease in reading performance. Sanders & McCormick (1993) approximate θ the ratio between text height (*h*) and reading distance (*D*) in arc minutes with:

$$\theta = 3438*h/D \qquad \text{Eq. 4}$$

For people with 20/20 visual acuity, the minimum readable text size (of the Latin alphabet) is five minutes of arc (Bailey & Lovie 1976). But the American National Standards Institute ANSI (1988) recommends a minimum size of 16 minutes of arc while the US military standard is 15 minutes of arc for the principal viewer and 10 minutes of arc at the maximum viewing distance (Musgrave 2001). Chapanis & Scarpa (1967) examined the effects of distance and size by comparing the readability of physical dials at different distances in lab experiments. The dials sizes and markings were proportional to the viewing distance, which ensured constant visual angles. The results showed that dials of the same visual angle were read more easily with increasing distance once the distance reached 72cm. The observed effect, however, was relatively small.

In terms of resolution fonts need to be at least five pixels high to be legible for standard ASCII fonts. The letter 'E', for example, needs three rows for the strokes and two for the spaces in between. Broadcasting companies have created guidelines for the use of text in TV content. The BBCi's design guideline states that the display of body text within the interactive television application's should not be smaller than 24 point and no text should be smaller than 18 point (Hansen 2005). But the guidelines fail to mention what the point size refers to when different size TV have different pixel sizes and in typography one point equals 0.353mm. The guidelines, furthermore, suggest that the presentation of text be in a sans serif font type (e.g. Gill Sans) using light colours on dark and that each screen should not contain more than 90 words of text. Contrast can suffer when text is encoded as part of a low bitrate video stream. Resizing TV content that includes text viewable at regular TV setups, e.g. from 720x576 pixels down to 120x90 pixels, can fall below five pixels or depending on the VR five minutes of arc, both of which renders text illegible. If not transmitted separately, text represents a medium within the medium of video. How perceived quality and legibility of text that is part of a video clip affects the QoE when the clip is present is currently unknown and is addressed by the study presented in Chapter 7.

2.3.6 Human visual temporal resolution

The temporal resolution of the eye is limited by the rods and cones on the retina, which respond to the accumulation of light over a certain period of time (Hart 1987). In order to perceive *apparent motion* e.g. individual flashes of a dot in different places - as a single moving dot, the time between the flashing dots must occur within a small enough time frame and within spatial limits on the retina. At a frequency of 100Hz apparent motion of a dot appears to be smooth and free of flicker. At very low frame rates (e.g.

when only one picture is shown every second), humans perceive these as individual images of dots. At adequate frame rates, this allows the human visual system to integrate the information and perceive smooth apparent motion, as opposed to true motion perceived in everyday life (Ramachandran & Anstis 1986). According to Steinman *et al.* (2000) the temporal resolution of human vision has been exploited since Roget's 1824 publication "*Persistence of Vision with Regard to moving objects*". See Figure 8 for an example of a Zoetrope and a flick book that presented short episodes of moving images content - early mediated visual information for entertainment. The human visual system cannot fully resolve objects whose images are moving too fast across the retina. Motion blur occurs and reduces the perceived spatial resolution of the moving object. Zettl (1973) defined three types of motion that can be present in moving images content. Primary motion happens within a shot type with a static camera, e.g. an actor is running across the screen. Secondary motion is due to the use of the camera. Through zooms, pans and dollies the frame is moving relative to what is depicted within it. Tertiary motion is defined as the sense of motion induced by editing different shots together.

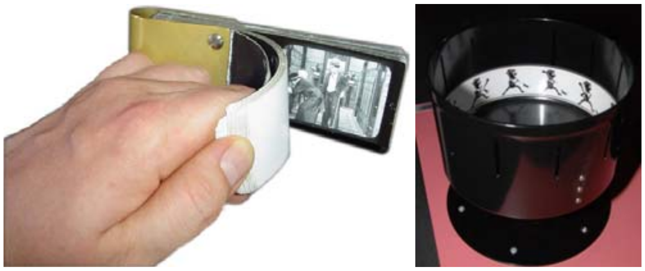


Figure 8: Flick book (left, photograph by Technische Sammlungen Dresden) and Zoetrope (right)

2.3.7 Visual field

Mathematically, the visual angle can assume any value between 0º and 180º and the viewing ratio between 0 and infinity for very small and very large objects. But the human visual system is limited in terms of the perception of very small objects by its visual acuity explained in Sec. 2.3.4 and in terms of very large objects by the size of the visual field. As objects become larger people need to move their eyes or heads in order to be able to see them in their entirety. Hatada *et al.* (1980) showed that - without head movement - the range that can be covered by eye-movement alone is 30º horizontally, 8º upward and 30º downward. Within this *effective visual field* people can process visual information effectively. The successively larger *induced* (20º -100º horizontally and 20º -85º vertically) and *supplementary visual field* (100º -200º horizontally and 85º -125º vertically) typically induce head movement or posture changes. In the *induced visual field* people can recognize the existence of visual stimuli. Abrupt stimuli in the supplementary field can cause a shift of a person's gaze. Apart from the different visual fields that are independent of viewing distance the eyeballs are limited by the range of possible orientations and the degree to which the lens of the eye can be deformed to focus on nearby objects (*cf.* Sec. 2.3.1).

2.3.8 Mediated reality

Moving images accompanied by sound can create a psychological illusion of a natural experience or of "*being there*" (Reeves & Nass 1998). How these depictions affect attention, memory, reflection and general evaluations of what is seen is subject to research. Criticism of TV generally concerns the limits of the medium as a whole in a societal context (McLuhan 1964), but not how its technical limitations affect human perception (which is the topic of this research). Depictions of scenes and objects on a screen can be seen as *symbols* that represent these objects. But children are willing to believe that people on the screen see them (Dorr 1980) or that they can manipulate objects depicted on screens by e.g. tilting the screen (Flavell *et al.* 1990). Although adults know better many reactions to depictions on screen are involuntary and provoke the same responses in the brain as real events would (Reeves *et al.* 1992). Depictions of faces receive more attention when they appear larger either through framing of the shot type or smaller viewing ratios (Reeves *et al.* 1992) - see the following section.

Hatada *et al.* discovered that it requires a picture with a 20 degree horizontal visual angle to induce a *sense of reality* for spectators, and that this effect became conspicuous at 30 degrees. Sense of reality was measured by the subjects' subjective judgments (on whether a white line on a black background was vertical). The line was shown before and after a stimulus of a horizontally rotated picture. Apart from this task, the authors obtained ratings on the induced *sense of reality* (in terms of e.g. feeling of depth and space fusion) using a 7-point ordinal scale. The results showed that pictures with horizontal visual angles larger than 20º to 30 º started to induce a sense of reality, and that the effect reached a plateau at 100º (horizontally) and 85º (vertically). The pictures were selected to give a feeling of expanse (an extreme long shot depicting a rural landscape) or of depth (a depiction of a suspended bridge with the vanishing point in the centre). Viewing distance had an effect on the *sense of reality*. At a constant angular size, larger viewing distances (3m) resulted in a greater subjective feeling of reality than shorter viewing distances according to Likert scale ratings (explained in Sec. 3.5). Shorter viewing distances required larger angular sizes of depictions to induce the effect. The authors advised against using viewing distances smaller than one meter to induce the sensation of reality. How these findings influenced the design of HDTV is explained in Sec. 2.5.2. Although the authors did not explain the procedure or the material used, they mentioned that their subjects tried to infer the distance between them and a depicted object, and preferred to view objects at their natural sizes - as implied by the distance of the camera that recorded the picture. The viewing distance at which a sensation of reality began depended on the visual angle of specific objects, rather than the visual angle of the whole picture. In depictions at visual horizontal angles smaller than 40º, objects appeared slightly larger to subjects than they really were. The size of captured objects depends on the distance between the camera and the object and the lens.

2.3.9 Shot types

The way in which objects are shot, edited, presented and decoded by the audience relies on established conventions (Thompson 1998). The different shot types used in film-making help the audience "read" the message the director wants to convey. The terms used to classify shot types can differ and popular usage of the terms deviates further. For consistency, the classification from Thompson (1998) will be used from this point on. This classification is centred on the depiction of people with the possible exception of the extreme long shot (XLS). In an XLS a person – if depicted at all - is barely visible and the recognition of the environment and/or the scene is more important (see Figure 9, left). In a very long shot (VLS) the majority of the frame is still concerned with the environment, in which the person is located. However, some details of the person such as clothing and gender are recognizable. In a long shot (LS) the person almost covers the frame from top to bottom (see Figure 9, right).

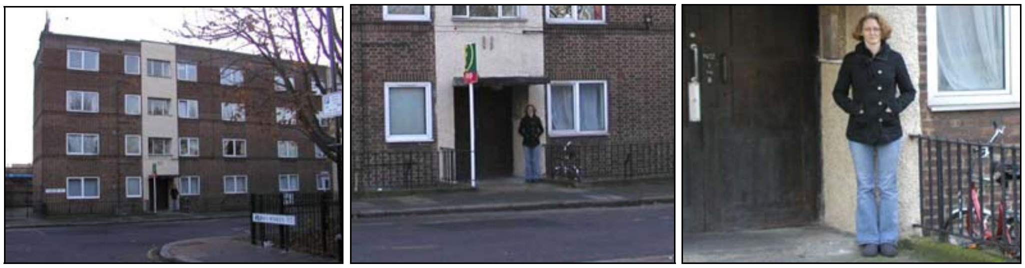


Figure 9: Extreme long shot (XLS), Very Long Shot (VLS) and Long Shot (LS)

In the medium shot (MS) the entire person does not fit into the frame anymore (see Figure 10, left). The eyes of the person can be seen clearly. The facial expression becomes predominant in the medium close-up (MCU) (see Figure 10, middle). The attention is drawn to the face and the background no longer important. On the close-up (see Figure 10, right) the attention is drawn to the person's eyes and mouth. Close-up shots induce a feeling of being pulled towards the screen (Hatada *et al.* 1980) and can convey threat and intimacy (Persson 1998). Each step from one to the next more detailed shot type represents a zoom factor between two and three. Shot types convey distance to people and might therefore convey social distance – explained in the following section - and its inherent emotional and social qualities.

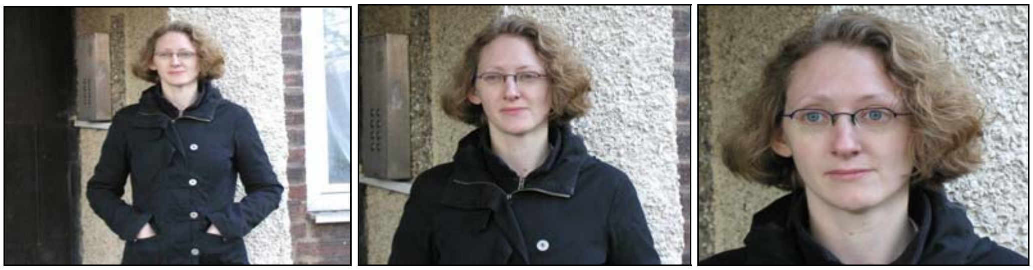


Figure 10: Medium Shot (MS), Medium close-up (MCU) and Close-Up (CU)

2.3.10 Social distance

The distance between an observer and an object limits what is visible in the visual field. However, when humans are involved distances have a social dimension because they depend on the relationships between the involved; norms on the social significance of distance vary between cultures. Hall (1966) classified distances between people into four groups:

1. *Intimate distance*: The interpersonal distance between zero and roughly 50cm is for people in intimate relationships such as couples and close friends. When one person enters this perimeter the other person may feel uncomfortable and/or find the first aggressive.
2. *Personal distance:* In daily life people usually assume a personal distance of at least 50cm to 1.2m to strangers.
3. *Social distance:* Face-to-face meetings and other business activities occur at a distance between 1.2m and 4m.
4. *Public distance:* Distances to people of more than 4m usually indicate no personal involvement with them. This distance is common for public speeches and performances.

Depending on these distances, a person takes up varying amounts of space in one's visual field and can be likened to what is depicted in a shot type. What one sees at an intimate distance equates roughly to what is shown in an MCU and CU (see Figure 10 middle and right). Personal distance provides an MS image illustrated in Figure 10 (left). Social distance affords a view between MS and LS depicted in Figure 10 (left) and Figure 9 (right). At public distances people appear somewhere between LS, VLS and XLS (Figure 9). At typical TV living room setups the angular size of the depicted person is typically much smaller than in the real life situation. But for example, testing a video communication system, Okada *et al.* (1994) noticed that some people felt pressured when depictions of a person on the screen became *larger than life-size* (*cf.* Sec. 2.6.7).

2.4 Capturing

Camera sensors are limited by their sensitivity to the light they need to capture visible objects -they need to collect light reflected from objects over time. Lenses are used to increase the amount of light that can be made available to the sensor: however, both lenses and sensors have limitations and non-linear responses across the spectrum of colour and light intensities. The research presented in this thesis does not address topics such as colour aberration and noise. Cameras can be chosen from a wide selection delivering different aspect ratios, contrast ranges, frame rates and resolutions.

2.4.1 Spatial resolution

Decisions about resolution are made at several points during the process of creation, editing, delivery and presentation of visual content. At the content creation stage, producers must decide, which resolution to target, and thereby select the equipment that can best capture content. The delivery of high-resolution content demands more resources; therefore, service providers need to find a trade-off between the added visual quality and the additional cost or reduction in the amount of content that can be delivered. As image resolution is reduced, there are fewer pixels to represent important information to the user. This may cause problems with some content types – such as sport – since fewer screen pixels are available to convey important details like the location of the ball when extreme long shots are used. According to the Nyquist-Shannon sampling theorem, when sampling a signal at a frequency *B* the highest frequency one can reliably reconstruct is smaller than *B/2* (Shannon 1949). A sensor with a spatial resolution of *B* picture elements (pixels) can resolve reliably a pattern with a frequency of a maximum of *B*/2. For television the resolving power of the camera was historically measured in line pairs (lp). The used criterion was when a black and white line grating of a certain width and contrast still appeared as distinct lines or became indistinguishable.

The relationship between the resolving power or resolution of a lens-sensor combination and the human eye at different contrast levels is typically not constant and can be described by the contrast sensitivity (CSF) or modulation transfer function (MTF) – see Figure 11. At high spatial resolution low modulation information becomes invisible. The highest sensitivity of the human eye is where the envelope of the visible lines peaks. Sensitivity changes with age - older people being less sensitive to higher frequencies (Owlsley *et al.* 1983). The relative sensitivity of the human visual system to moving image depends furthermore 1) on the speed at which depicted objects move and 2) the temporal frequency (Kelly 1979). At high frequencies, contrast-sensitivity plateaus for low resolutions and drops off for higher resolutions (Fujio 1985). Any real-life scene has theoretically an infinite resolution but the resulting resolution achieved with a sensor of *B* pixels in height or width is between B and B/2. Objects in the recorded scene with spatial frequencies higher than half the sensor's spatial resolution might incur artefacts known as *aliasing*: existing high frequency patterns above the human acuity threshold might become visible and patterns that did not exist in the original scene might manifest themselves in the sampled material, for example, moirés (Ware 2000). In TV production the ties of news anchors with very detailed patterns resulted in – for the viewers visually irritating - moirés. In television reproduction the Kell factor – the ratio between the achieved resolution in line pairs and the line pairs of a camera - was originally assumed to be 0.64. This means that television footage, which was progressively captured with a camera with a 100lp resolution, would have a resolution of 64lp (Kell *et al.* 1934). Researchers suggested various values (see Table 2). Much later, Martin (1985) used a factor of around 0.5 for HDTV requirements. Robin (2003) hypothesized that these different values were due to *"differences in the picture display systems used by different observers, as well as subjective picture quality appreciation"*. Poynton (2003) dismissed the Kell factor due to the subjective nature of the tests on which they were based. He proposed a factor of 0.7 for interlaced TV, which includes both the effect of the sampling and interlacing (see Sec. 2.4.3) but he does not cite any publications that provide empirical evidence. Overall, the literature on the spatial resolution of moving images seems inconclusive about the achieved resolution due to different measuring approaches, camera/display combinations and the contribution of the content in terms of the depicted moving detail.

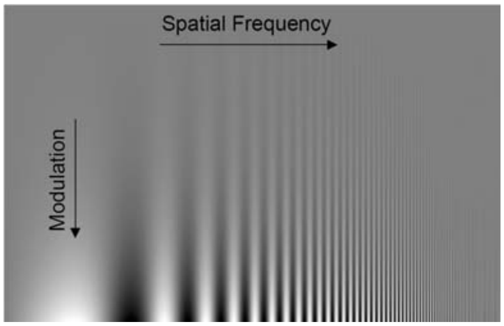


Figure 11: Modulation Transfer Function

Table 2: Achieved resolution factors

Factor	*proposed by*
0.53	Mertz & Gray (1934)
0.71	Wheeler & Loughren (1938)
0.82	Wilson (1938)
0.85	Kell, *et al.* (1940)
0.70	Baldwin, Jr. (1940)

2.4.2 Temporal resolution

The number of pictures per second in a video clip has to be high enough to induce apparent motion when viewed by a human through a display: motion appears choppy or jerky when captured at low frame rates. Lower frame rates have been used for animated content to cut the cost of producing more pictures (pictures were repeated instead). Slow-motion cameras record at higher frame rates to be able to replay a sequence over a longer time period at standard frame rates. Insufficient temporal resolution of a camera can result in noticeable artefacts such as the *reversing wagon wheel* effect (Ware 2000). However, this effect can still happen at the display stage even at high frames rates due to the limits of human visual perception. The analogue TV delivery chain was designed around the requirement of portraying movement naturally. Reducing frame rates has been a popular measure in computer-based video to reduce the required volume of information requiring processing, storage and transport through the network. According to Hellström (1997) spelling sign language requires 25 fps to capture all letters in at least one frame. However, more recent studies have shown that the spatial resolution of the face carrying meaningful cues for sign language appears to be more important than the spatial resolution of the hands (Muir & Richardson 2002), (Agrafiotis *et al.* 2003). Many TV and video cameras do not record pictures progressively but in an interlaced way. The camera alternates reading out the sensor's two fields – the even and the odd lines of recording pixels. Television content captured at a nominal 25 fps actually consists of 50 fields per second.

2.4.3 Spatio-temporal trade-off

For static scenes, interlaced scanning results in the same vertical resolution as a progressive scan but it requires only half the bandwidth to capture and transmit because only 50 half-pictures are being captured per second in comparison to 25 full frame progressive scanning. Due to the higher temporal refresh rate interlacing improves the portrayal of motion (Poynton 2003). Interlacing introduces two artefacts - line twitter in the captured material especially for non-static scenes, and - depending on the presentation device - line flicker. Interlaced scanning reduces the vertical spatial resolution of captured dynamic scenes by an interlace factor of 0.7 according to Wood (2004)[2]. Section 2.5.2 on SDTV provides more details on interlacing. In digital capture environments the technique of *motion blurring* can reduce jerky motion caused by too low temporal resolution - at the expense of spatial resolution. The technique uses longer exposure times of the camera sensor or by interpolating two or more frames into a single frame. This reduces the jerkiness of moving objects, but makes them appear more blurred.

2.4.4 Digital quantization

As a digital camera samples a spatial scene over time with a sensor, the question remains how the camera quantifies the amount of light that each pixel collects. When digitally quantifying the amount of light captured by a pixel, the number of discrete possible target values is an important criterion. Possible quantization levels range from coarse binary black and white over gray scales and for colour up to *true colour* quantization. The amount of storage required to store a quantized value grows logarithmically with the target range of quantifiable values. Along with the spatial and temporal resolution, quantization determines the required space of the data for storage and delivery and affects the MTF.

2.4.5 Framing

Since TV receivers are not perfectly aligned, TV content is shot in a way that ensures that all viewers will have their screen filled by a picture. The European Broadcast Union (EBU) suggests that television programme makers frame pictures so that all action is contained in the *action safe area* and all graphics in the *graphics safe area*. For 16:9 SDTV, this means that all graphics should be displayed in the central 516 lines and 562 horizontal pixels and all essential action should take place in the central 536 lines and 652 pixels (instead of 576 lines and 702 horizontal pixels) of the recording camera. Depictions outside these areas cannot be guaranteed to be seen by all receivers (EBU 2008). In (Tanton & Stone 1989) the entire captured picture is referred to as the *underscan.*

2.5 Representation and delivery

The large amount of information necessary to represent moving images can be stored in uncompressed digital formats or in compressed formats. The originally captured information cannot be fully recreated

[2] Relying on Poynton's (2003) recommendations I assume a combined Kell and interlace factor of 0.7 for all calculations of the resolution of footage captured from interlaced TV.

when lossy compression is used - but depending on the format and compression ratio the difference between the lossless and compressed versions can be small. Video encoders that compute the digital representation of a certain format typically consider quantization as a parameter that affects the spatial clarity of the encoded image as does the number of pixels and the frame rate. Depending on the amount of compression the original resolution is further reduced both in the spatial and the temporal domain.

One important ramification of digital mediation is the added encoding delay, compared to an analogue signal that could be transmitted right after it has been read out from the camera. To achieve higher compression, the encoder might need to wait for the capture of future frames to perform *bi-directional* compression (Haskell *et al.* 1997). Computing the compression itself requires time. Unlike analogue TV sets a digital receiver does not simply change the frequency to tune into another TV channel – it has to wait until it can reconstruct and present a full frame. This increases the time required to switch channels - an important criterion for the user experience (see Sec. 4.3.4).

The way the content is delivered has a major effect on the possible uses of a mobile TV service. Unlike stationary systems mobile receivers are capable of receiving data while moving within a certain speed range. But signal interference and coverage outages might result in erroneous reception and loss of data and affect the QoE. When transmitted through broadcast or IP-based networks, the loss of parts of the information can distort the picture through a range of visual artefacts.

2.5.1 Digital format

Video data can be represented in various formats. Many digital TV broadcast standards and DVDs rely on the Motion Picture Expert Group 2 (MPEG2) standard (Haskell *et al.* 1997). This format makes use of spatio-temporal redundancy in adjacent frames and is the basis of many of today's digital television standards (*cf.* Sec. 2.5.2). Intra-coded (I) frames are stored and transmitted in full but predictive and bi-directional frames encode only the differences to previous and following frames. A broadcast TV channel in MPEG2 requires less spectrum and therefore digital TV can offer a larger number of channels. Digital distribution allows for different qualities for each channel or individual programme.

2.5.2 Standard definition television (SDTV)

Originally aimed at providing a vehicle for society to disseminate information, TV services today mainly target entertainment and mediating experiences (see Sec. 4.3.3). In Europe, the soon-to-be phased-out analogue SDTV was based on the original inception of television in the 1930s. Television receivers were designed around the cathode ray tube (CRT) which involved a vacuum tube with an electron gun at one end and a fluorescent plane on the other end. When TV was invented, the production constraints of the vacuum tube favoured an aspect ratio that was closer to a square than a wide rectangle (Schubin 2007). The fluorescent plane has a structure of lines, parts of which light up on the impact of electrons from the gun. Broadcast television in Europe uses a 25 fps refresh rate with an interlace ratio of 2:1 resulting in the presentation of 50 fields or scans per second. Each full frame is made up of two fields for a full raster of 576 lines (720x576). Interlacing cleverly kills two birds with one stone by trading off spatial resolution in moving images for greater temporal resolution. The doubling of the temporal resolution reduces artefacts of moving objects – albeit at a lower spatial resolution. Still, interlaced scenes benefit from a higher spatial resolution than a 50fps progressive scan with half the lines would achieve.

There are different standards to convey colour information, e.g. PAL and SECAM (for more detail see (Poynton 2003)). Mitsuhashi (1991) equated the picture quality of SDTV to roughly that of 16mm film but comparisons are complicated by interlacing and interline twitter artefacts. Hatada *et al.* (1980) claimed that watching standard resolution TV "*is tiring to look at because it creates a condition of a fixed semi-stare,*" but they did not back this up with results or references. In the design process of digital and analogue TV, researchers used video quality assessment techniques to obtain feedback from viewers. Their approaches included psychophysical scaling (see Sec. 3.5.1), and employed indirect indicators such as preferred viewing distance (PVD), see e.g. (Jesty 1958), (Lund 1993), (Ardito *et al.* 1996) as operationalizations for visual quality. Mitsuhashi & Yuyuma (1991) identify the coarse line structure, the low resolution and the limited sensation of reality due to its limited visual angle of around 10º horizontally as the major downsides of SDTV. However, their argument might - at least partly - have been based on assumptions about smaller viewing distances in Japanese households – between 2.66m (Kubota *et al.* 2006) and 2.5m (Fujine *et al.* 2008) – than in Europe (3m) and the US (10ft).

2.5.3 High-definition TV (HDTV)

Since 1925 high-definition television has been referred to as the next technical advance in comparison to the standard that had been in place (Schubin 2003). What people currently refer to as HDTV is based on research carried out in Japan in the 1970s and 80s. However, there is a whole range of standards that are labelled HDTV - among them are 1152x720 (one Megapixel) and 1920x1080 (two Megapixels). Unfortunately, a number of early studies on HDTV exist only in Japanese, and are therefore not covered in this review. The motivation behind HDTV was to create a different viewing experience - the larger image was to induce a sense of reality, according to the findings on the human visual field of Hatada *et al.* (see Sec. 0). The idea was to provide a larger picture with the same angular resolution as standard definition TV (Poynton 2003) that subtended a 30º visual angle horizontally. At the same time, HDTV provides more visual detail, i.e., definition of depicted objects.

People show an overwhelming preference for widescreen 16:9 (*cf.* Sec. 2.6.1) over the SDTV 4:3 format, and at that aspect ratio, a 30º horizontal visual angle can be achieved with a VR of 3H. The literature often refers to the corresponding viewing distance – three times the picture height – as HDTV's design viewing distance (dvd). At the typical TV viewing distance of around 3m - the benefits of HDTV can only be enjoyed on relatively big screens. To render the line structure invisible at this viewing ratio, around 1100 lines are required. This, in short, was the idea of the engineers involved in the design of HDTV: to convey the new experience at the same angular resolution as analogue SDTV. HDTV required roughly more than four times the traditional bandwidth.

2.5.4 Digital TV

The introduction of mobile TV overlaps with the switch-over from analogue to digital TV. In Europe there are standards for digital terrestrial (DVB-T), satellite (DVB-S) and cable (DVB-C) broadcast. The main reason for the switch-over is more efficient usage of the spectrum that has been allocated to analogue TV. The switch from analogue to digital can only be made if the current base of TV receivers can continue to be used. Inexpensive set top boxes (STB) turn the digital broadcasts into the standard signals that analogue TV receivers can display. Similarly, Internet service providers have started to provide content over the Internet protocol to TVs (IPTV) via STBs.

In analogue transmission, the frequency bandwidth imposed a fairly uniform limit on the spatio-temporal resolution of the content. Interlacing addressed to some degree the problem between static and dynamic scenes. Static scenes would benefit from the larger amount of lines and dynamic scenes from the higher temporal update frequency at the expense of spatial resolution. Apart from the differences between more and less dynamic scenes, the bandwidth of all programs and channels were identical. Digital distribution in turn allows for a large diversity of quality profiles. Each channel, programme or individual piece of content can be encoded individually with a unique quality profile based on economic considerations. Some content types lend themselves to better compression than others. Typically this profile is a limit on the encoding bitrate that can be used for the content. This envelope makes it easier to think of the content in terms of blocking a certain amount of bandwidth over time. However, with a fixed encoding bitrate, the video quality of a programme will not usually be uniform: scenes with a lot of motion, or scene changes, require more encoding bitrates than those with few temporal changes. The latter will appear to be of higher quality.

2.5.5 Computer-based TV

For broadcast television, the display of audio-visual information was approached from a conservative point of view - to deliver the most natural rendition of a mediated window of the world that was technically feasible. In digital environments, the presentation of information is independent of the capture and transmission process; thus, a number of studies have tried to find feasible parameter settings that are below human audio-visual perception thresholds, but are still fit for purpose in the context of a given task or the enjoyment of multimedia content. This was especially the case in digitally-mediated environments like desktop video conferencing in which the transmitted visual information was not directly modulating a cathode ray tube. Since the availability of broadband Internet computer screens have become the *third screen* – after cinema and television - to render moving images for entertainment purposes. Unlike broadcast TV, most content delivered over the Internet is delivered to a single receiver at a time. Because of individual data delivery, Internet-based services have tried to find configurations of services that minimise the amount of data that needs to be transferred. Computers are increasingly used to provide services on top of existing broadcast offerings. Zattoo delivers regular broadcast TV content in real-time as a streaming service. The service offered through Slingbox allows for reception of broadcast TV content, which is then streamed through the Internet to the user who can be located outside the country. Video on demand (VoD) services allow service providers to deliver content to single household at a time of the consumers choosing.

YouTube lets people share their content and consume available clips on demand. Furthermore, it offers searching for clips through text queries, rating, recommending and annotating the hosted material. Users can ‘charge’ their iPod video players with content downloaded from the iTunes store servers and then consume them at any time - from once to as often as desired, depending on the charging model. Servers and network bandwidth of on-demand content delivery services have to be dimensioned to satisfy peak demands making them expensive to operate. Peer-to-peer delivery, which is used in services like e.g. Joost (2009), scales better for popular material since many people are watching and thereby sharing, i.e., re-distributing popular content. This removes the distributors' servers and their bandwidth as the bottleneck in content delivery but at the same time cannot guarantee the general availability of a given content item.

2.6 Displays

Viewing mediated visual information requires a display that can recreate moving images. Visual information can be shown on a light-emitting display or through projectors onto a reflecting surface or directly onto the retina. The two most important display types for the work presented in this thesis are pixel-based and line-based displays. The former represents the current standard for mobile digital devices and the latter has been the major standard for TV systems since its inception. No matter which technology is used for mobile TV screens they should have high contrast, high luminance and a high viewing angle to support viewing under different circumstances and by multiple viewers.

2.6.1 Aspect ratio

The aspect ratio of an image describes the ratio between its width and height. Different aspect ratios have been adopted over time. Those in which the width is larger than the height are referred to as *landscape* - the converse is called *portrait*. In the early days of filmmaking standards had not emerged yet. Russian director Sergej Eisenstein suggested a square screen (Dancyger 2008) on which footage of both portrait and landscape format could be displayed depending on which best suited the depicted scene. However, all current standards like SDTV (4:3, 1.33:1), widescreen SDTV and HDTV (16:9; 1.78:1), and cinema film (1.85:1 to 2.39:1) use landscape formats. This might be due to the human field of view being wide rather than high but Schubin (2007) argued that this choice is mainly due to gravity, which makes more movement happen in horizontal than in vertical directions. When given the choice people prefer widescreen (16:9) over standards 4:3 aspect ratio for TV content (Pitts & Hurst 1989). This holds true when both formats are presented at equal height, equal diagonal, equal area and equal width (albeit to a lesser degree) irrespective of viewing distance and screen size. Modern TVs, projectors and mobile devices can adapt aspect ratios. Depending on what other information needs to be displayed on mobile devices size and/or aspect ratio might need to be adapted. I will cover content adaptation in more detail in Sec. 4.6.3.

2.6.2 Aspect ratio adaptation

Content producers capture video in a certain aspect ratio and resolution. Since the aspect ratios of SDTV, widescreen TV (16:19) and cinema differ captured material has been adapted to different displays. When content producers oppose altering the produced footage letterboxing and pillarboxing (Poynton 2003) are used to adapt to other aspect ratios. In letterboxing black bars are added above and below the footage to adapt wider aspect ratios (e.g. cinema 1.85:1) to narrower aspect ratios (e.g. SDTVs 4:3). In the reverse case these black bars are added on the left and right of the footage e.g. to show 4:3 TV content on 16:9 screens. This approach does not make use of the whole screen. If distributors want to fill all of a non-matching target display, they can crop or stretch the material.

Cropping superimposes a window of the correct target aspect ratio on top of the footage and leaves out all of the visual footage that is not contained in that window. In theory this can be done with a static window that centres on the middle of the original footage. In practice, however, especially in the adaptation of cinema to narrower aspect ratios, the *pan-and-scan* approach is used. Historically pan-and-scan required a human operator to film the projected cinema footage with a TV aspect ratio camera, and choose which part of the picture to show and what to omit. In stretching content, the depicted objects are changed from their original form. This can be applied to the whole picture either in the horizontal direction when adapting narrower to wider aspect ratios or vertically for the opposite adaptation. Some widescreen TV sets do not stretch the whole picture evenly but take slices of the left and right side of the picture and stretch them (Söhne *et al.* 1998). This *panorama vision* results in visible artefacts, especially in camera pans or horizontally moving objects or text but it leaves the centre of the screen undistorted.

2.6.3 Spatial resolution

At the presentation stage the capabilities of the presentation devices determines the resolution at which content can be presented. The number of lines multiplied with the resolution per line or the number of addressable picture elements (pixels) resolution of a display is a core parameter also known as addressability (Daintith 2004). Along with the temporal resolution it defined the required bandwidth for the transmission of analogue SDTV. Although SDTV sets come in a range of sizes the number of lines is constant and depends on the standard used. Television sets have a structure of stacked lines that are used to display the incoming signal. For larger SDTV sets these lines are simply larger than they are for smaller sets. Research by Thompson (1957) on the influence of the horizontal TV scan lines (the vertical raster) showed that people preferred to reduce their viewing distance from the screen when the line structure was not visible. This suggested that line visibility needed to be considered as a potential factor in peoples preferred viewing distance (PVD). Sproson's (1958) results showed that with larger viewing distances of 4, 6 and 8H line visibility decreased but that the assessment of 405-line monochrome TV pictures was unaffected by the visibility of the lines. The spatial resolution of a display is dependent on contrast. Lowering the contrast affects the modulation transfer function (MTF) and high frequency detail is lost. Ardito *et al.* showed that HDTV is reduced to perceived standard definition television resolution at low contrast levels, e.g. when the luminance of the screen was not sufficient (Ardito et al. 1996).

Displays that are not tied to the presentation of TV signals are less restricted in the grid of picture elements and content of different resolution can be presented on them either natively or scaled (for more on content adaptation see Sec.4.5.4). Computer screens come in a huge variety of resolutions and aspect ratios. Mobile displays currently range from VGA PDAs (480x640 pixels) and high end 3G or DVB-H enabled phones (320x240) to more compact models with QCIF resolution (176x144). Other mobile devices such as DVD players and computer laptops have even higher resolutions. The pixel density on devices is typically expressed in pixels per inch (ppi). Typical mobile devices range between 80ppi and 200ppi. The effects of up-scaling or stretching broadcast content to a screen with a higher resolution as common in currently available *HD-ready* TV sets are the target of proprietary research. Philips uses a non-linear up-scaling method called *Mobile PixelPlus* to fill a screen of higher addressability (Zhao *et al.* 2007). Reduction of resolution decreases the amount of data required to represent the images, allowing the transmission of video over constrained bandwidth networks. The exact boundaries within which up-scaling does not impact the video quality are unknown but low resolution (160x120) video reduced students' satisfaction with a distance learning application when compared to 320x240 resolution (Kies *et al.* 1996) in a computer desktop setting.

2.6.4 Spatial resolution’s effect on shot types

The perceived value of resolution depends on what is depicted. People judge the sharpness of photographs depicting a landscape differently from the portrait of a person. Frieser & Biedermann (1963) noted that *“for a portrait an unusually low MTF was sufficient for an impression of good sharpness”.* A landscape required a higher resolution than a portrait to be considered of good quality. Kingslake (1963) attributed this to the small size of distant objects for which blurring can be detected more easily than for near objects. Corey *et al.* (1983) found a matching systematic relation between the subjective picture quality and the distance of the subject to the lens (or the magnification). With increasing distance the perceived quality decreased. These results suggest that when presented at low resolution closer up shots will be perceived to be of higher quality than longer distance shots. Very high resolution, however, might not result in the highest quality. For a portrait the highest resolution in terms of the modulation-transfer function (MTF) was judged of worse quality than a slightly lower quality in a study by Frieser *et al.* (1963).

High spatial resolution can yield higher utility if it supports specific tasks that are important to users. If viewers of a football video want to identify individual players, not being able to do so will affect the perceived visual quality (McCarthy *et al.* 2004a). Research in face recognition has shown that human observers require at least 15 pixels per face (in vertical resolution) to be able to identify them (Bachmann 1991); (Bathia *et al.* 1995). In a study by Barber & Laws (1994) a reduction in image resolution (from 256x256 to 128x128) at a constant image size led to a loss in accuracy of emotion detection, especially in a full body view, which is equivalent to a long shot (LS). For this task resolution was the most important factor over frame rate, grey levels, image size, display addressability and pixel size of the display.

2.6.5 Temporal resolution

The display of moving images relies on picture frames, which are presented in rapid succession and induce apparent motion in human viewers. Limitations in display device technology made flicker between pictures or the lines that make up the picture a problem because it degraded the viewing experience. Flicker is exacerbated in large displays that span the human visual field and subtend the peripheral vision in which humans are more sensitive to motion and therefore flicker. Flicker is further increased by ambient lighting (Poynton 2003). In terms of temporal resolution displays have different refresh rates, also called flash rates. CRT displays typically do not emit light continuously. At any given time parts of the screen are black. In TV sets the refresh rates of the screen are between 50 and 60Hz. At this rate the flicker is hardly noticeable to the human visual system. Flicker is not a concern in displays in which pixels are constantly emitting light e.g. in liquid crystal displays (LCD). However, the speed with which a pixel can be changed from one luminance state to another constitutes a limiting factor for the depiction of moving pictures on those displays. The proprietary *Natural Motion* approach by Philips supposedly reduces the jerkiness of low-frame rate content by generating intermediate frames from the set of frames at the receiver side (de Vries 2006) but there are no studies freely available that back up this claim but it should be similar to motion blurring. Apteker *et al.* (1994) assessed the *watchability* of various types of video at different frame rates (30, 15, 10, 5 fps). They manipulated the static/dynamic nature of the video and the importance of the video and audio information to understanding the message. Dynamic videos were characterized by scenery like sports footage, and static video images were characterized by scenes like talk shows. Compared to a benchmark of 100% at 30fps, video clips high in visual importance dropped to a range of 43% to 64% watchability when displayed at 5 fps, depending upon the importance of audio for the comprehension of the content and the static/dynamic nature of the video. Overall, watchability of the video clips dropped significantly with each 5 fps decrease in frame rate for 15, 10 and 5 fps. The authors concluded that video degrades in perceived quality with decreasing frame rates but that the ratings are highly contingent upon factors such as the importance of the audio or video information to message comprehension as well as the temporal nature of the imagery.

Although a number of studies have shown five fps to be a feasible minimum frame rate for video-conferencing (Tang & Isaacs 1993) and tasks such as monitoring rats (Chuang & Haines 1993) or intellectual tasks (Masoodian *et al.* 1995), the lower bound for an enjoyable video experience on mobile devices was shown to be higher. Participants who saw football clips on mobile devices found the video quality of football content less acceptable when the frame rate dropped below 12fps (McCarthy *et al.* 2004a). Comparable displays on desktop computers maintained high acceptability for frame rates as low as 6fps. The reason for this seemingly higher sensitivity to low frame rates on mobile devices is not yet fully understood, but highlights the importance of measuring video quality in realistic setups.

For a 30 fps video the window of synchronization with the audio track is ±80ms (Steinmetz 1996). People can not reliably detect deviations smaller than 80ms. Synchronous playback of sound and video affects the overall audio-visual quality but the temporal window of synchronisation depends on the video frame rate (Knoche *et al.* 2005), (Vatakis & Spence 2006).

2.6.6 Viewing distance

A change in viewing distance changes the amount of light that reaches the retina from the screen. Increasing the viewing distance lets more people share a screen and results in smaller images with lower contrast ratio. But at the same time it increases the visual quality as the resulting angular resolution of the depicted content increases. At some points the individual TV lines and pixels become invisible. Looking at the viewing distance in isolation is prone to oversimplification as many other variables change with it. However, there are a number of findings that highlight that viewing distance depends on factors that are not confounded with angular resolution and size. In Sec. 2.3.1 I described limitations on the practical range of viewing distances for displays due to accommodation and vergence of the human visual apparatus. Based on this we can assume that viewing distances (in terms of the focal point of accommodation[3]) for any device will be larger than 15cm regardless of the size and visual resolution of the display.

Most of the research on viewing distances in relation to screens has been conducted in the field of television engineering. Since there is no comparable body of literature in the mobile multimedia field I will review the previous research on viewing distances of SDTV and HDTV. In many cases these results consider viewing distance in relation to screen height and the resulting preferred viewing ratios might apply to the mobile domain. For example, extrapolating from their work on very large scale displays Sadashigo and Ando suggested that the viewing distance should be 53cm for a personal display of 6.1cm

[3] Near-eye displays are operated at very close distance but their focal point lies further away.

in height (Sadashige & Ando 1984). I revisit the influence of viewing distance in relation to angular resolution and size in more detail in Chapter 11.

Field studies

Fink (1955) found that the average of chosen viewing distance was 3.3*H* in cinemas and 7.1*H* for 480 line TV. Kaufman & Christensen (1987) reported similar values between 2*H* and 4*H* in large theatres - relatively extensive spaces with almost no ambient lighting. Nathan *et al.* (1985) found that larger displays correlated with greater viewing distances. In their study, viewing distance of regular TV varied also with the age of the viewers. The average viewing distance for 17 year olds and younger was 2.25m (VR 7.8*H*) whereas adults watched from 3.37m (11.7*H*). The average screen height in their study was 29cm. The study failed to address whether screen size and room size were correlated. Thus it cannot be ruled out that people with larger TV screens also had bigger living rooms and this correlation was responsible for the correlation between screen size and viewing distance. The study did not explain the difference between age groups but it did report that children were more mobile than adults and much less likely to sit or lie on furniture while watching TV. This provides some evidence that the layout of furniture has a strong effect on viewing distance. Recall from Sec. 0 that most research on viewing distances in the home as the Lechner and Jackson distance as well as Tanton & Stone's recent study with larger average screen sizes all indicated a fixed viewing distance in the home at around nine feet or approximately three meters.

Lab studies

Without the constraints common in people's homes, lab studies found different results on people's preferred viewing distances (PVD). In a series of five studies Lund showed that participants' preferred viewing ratio was not a constant 7*H*. With increasing image size, and independent of resolution, the preferred viewing ratio approached 3*H* or 4*H*. With decreasing image sizes the ratio approached 7*H.* This tendency was true for both passive consumption preferences of naïve participants as well as for expert assessors who were asked to assess the video quality.

Ardito found that the PVDs for moving images was further away than for still pictures. For HDTV content Ardito he predicted a viewing distance *D* (expressed in picture heights *H*) of $D=(3.55\ H+90)/H$. Although he did not test small mobile screens he interpolated from a range of HDTV screen heights of 198cm to 15cm that for screens with a screen height close to zero the viewing distance would be 90cm. In order for people to choose HDTV's design viewing distance of 3H Ardito *et al.* had to present HDTV video in a darkened room with a projector on a screen with a diagonal of 4m (a screen height of 2m).

For subjective video quality assessment methods for multimedia applications the ITU suggests viewing ratios between one and eight in their recommendation series P.910 (ITU-T 1999). But for subjective video quality assessment of TV material the ITU specifies PVDs depending on the screen height (see Table 3) in its recommendation BT.500-11 from 2004. The recommendation contains a graph that illustrates the relationship between screen height and PVD for screen sizes between 18*cm* and 2*m*. A power function $f(x)=76.5\ x^{-0.41}$ describes the relationship of screen height in *mm* to PVD (*cf.* Table 3) reasonably well ($R^2 = 0.97$). According to ITU these values should be applied for both SD- and HDTV *'as very little difference was found'* between the two. Screen heights smaller than 18cm and smaller resolutions are not covered by the ITU's recommendations but considering the trend their PVD should be 11*H* and higher.

Unfortunately, the ITU does not specify the source or study of these recommendations and they can therefore only be interpreted as the viewing distance that the ITU prefers i.e. recommends for subjective viewing tests (ITU-R 2004). However, as can be seen in Figure 12 the ITU's relationship between screen height and PVD follows the trends suggested by Lund and Ardito's results.

Table 3: PVDs depending on screen height recommended by the ITU in rec. BT.500

Screen diagonal (inches)		*Screen height (H)*	*PVD*	*PVD*
4:3 ratio	*16:9 ratio*	*(m)*	*in screen heights (H)*	*in m*
12	15	0.18	9	1.7
15	18	0.23	8	1.8
20	24	0.30	7	2.1
29	36	0.45	6	2.7
60	73	0.91	5	4.5
>100	>120	>1.53	3-4	>4.6

In summary, the viewing distance at which people choose to follow audio-visual content under lab conditions is not fully understood and the size of the screen, the viewing conditions such as ambient lighting and the resolution of the content influence this decision. Surprisingly, the ITU's recommendations on viewing ratios for subjective TV video quality disregard the field research findings in which people kept a viewing distance of approximately three meters in the home and did not change it in response to screen size. Yu *et al.* found no statistical difference in assessors judging video quality impairments when the NTSC SDTV material of 525 lines and 30 fps was presented at 3*H* or 5*H* (Yu *et al.* 2002). But no research has shown that this holds for VRs larger than 5 and up to 9 which would justify the suggested PVDs in ITU recommendation BT.500 when quality impairments are measured.

Sugama *et al.* (2005) found that for still pictures of identical angular resolution of 27ppd on a 100ppi monitor – all shown at 6*H* - the subjective video quality was higher when they were viewed at a close distance of 40cm in comparison to viewing distances of 80cm and 1.6m (Sugama et al. 2005). However, the study did not control for the addressability of the display. At the closest distance the angular resolution of the display was 27ppd but for the largest viewing distance (1.6m) with the medium (54ppd) and large (100ppd) images the pixels were close to and above the human discrimination threshold.

Figure 12 displays the results of the most prominent studies described above. It plots the resulting viewing ratios for screens of different heights based on the viewers' PVDs. Values that were obtained in dark rooms are marked with underlying shadows. These results span a range of display techniques (computer CRT screens, TV sets, rear-projectors, projectors) and locations (living rooms, small labs, meeting and larger rooms) and lighting conditions. People have opted for viewing ratios smaller than 5 only in darkened rooms of considerable size. It is unclear, how much these preferences depended on 1) the similarity of the rooms to either cinema setups or living rooms, 2) resolution changes to higher contrast in darkened rooms, 3) the screen size, or 4) screen addressability.

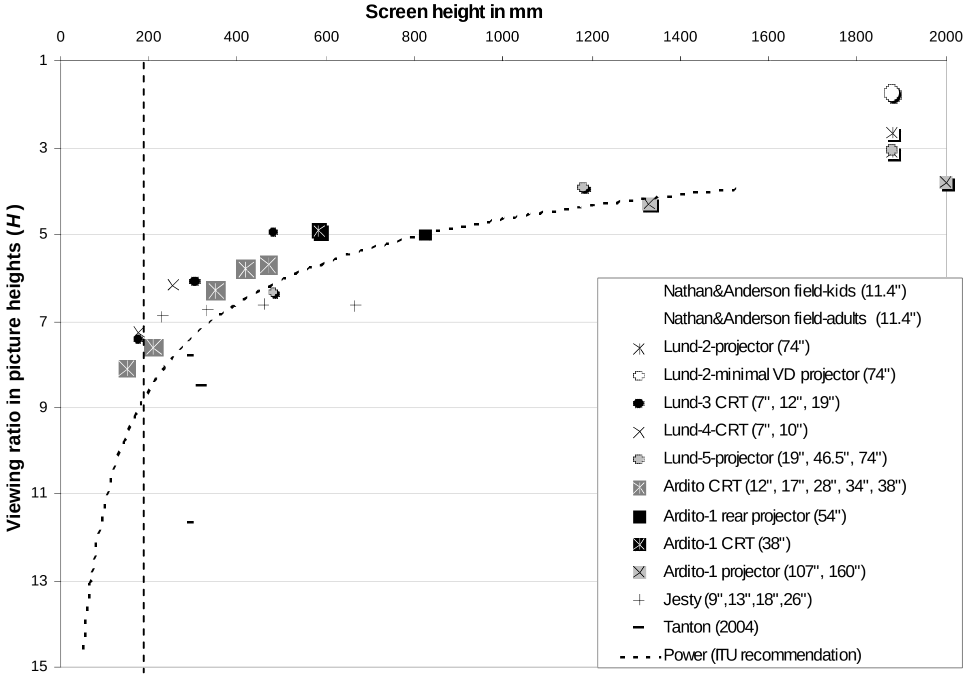


Figure 12: Preferred viewing ratio in relation to screen height – including the ITU rec. BT.500 and its limit at 180mm (vertical dashed line)

2.6.7 Viewing ratio and angular size

Research on the effects of angular image size was largely grounded on tasks dating back to studies by e.g. Steedman and Baker (1960), which were based on still pictures. However, many of these studies concentrated on human thresholds, e.g. in detection, recognition and identification tasks in a military context, e.g. (Johnson 1958), concerning the resolution of objects. But a number of studies have provided evidence that angular size affects the viewing experience in various ways.

In video conferencing larger video displays prompted people to view the screen from proportionally larger distances (Duncanson & Williams 1973), (Lombard *et al.* 1996). Hatada *et al.* (1980) found that the angular size of the display was not sufficient for describing the effect of display size but that the absolute picture size or the absolute viewing distance needed to be considered.

Viewing angles larger than 20º horizontally induce a *sense of reality* (*cf.* Sec. 0). To achieve a 30º visual angle with a 16:9 aspect ratio the viewing distance needs to be 3*H* (Mitsuhashi & Yuyama 1991). Introducing screens of this size into people's living rooms was not without problems. In 1989 the BBC (Tanton & Stone) conducted tests in homes using 16:9 photo prints of a landscape – akin to an XLS - of different sizes to assess people's desirable, optimum and practical screen sizes in the living room. Without a rearrangement of furniture, the mean optimum viewing ratio was only 6.1*H*. Participants prepared to rearrange their living room to accommodate HDTV had a slightly smaller 5.6*H* optimum viewing distance. Based on their results, the authors expected viewing distances between 4*H* and 6*H.* A number of participants remarked that *"they would not relish watching a 'talking-heads' interview scene on such a large screen from such a close distance".* This matches the findings of Okada *et al.* (1994). The authors softened this criticism by pointing out that production would consider different framing for

HDTV content, and that problems with larger-than-life images *"could be alleviated in the receiver by the provision of a switchable 'under-scan' facility"*. Although a switch to under-scan, in which the whole picture is shown, would make the picture slightly smaller the overall order of size due to the used shot type would not change much. The studies by Lund and Ardito *et al.* (see Figure 12) have shown that people choose viewing ratios in the range of 3*H* for HDTV only in large dark rooms and with very large projections.

According to Reeves *et al.* (1999) pictures subtending a larger visual angle make for a better viewing experience (34º or 2.2H in comparison to 17º or 4.4H) in terms of e.g. excitement, feeling of involvement in the action and realism. But they found no difference in arousal and attention between watching content on 2" and 13" screens at VR of 12 and 6.5 respectively. Results from studies on TV pictures by Reeves & Nass (1998) and Lombard *et al.* (2000) found that larger image sizes are generally preferred to smaller ones - see also (Lombard *et al.* 1996). At constant angular resolution larger images are perceived to be of higher quality (Westerink & Roufs 1989). Visual fatigue in Japanese viewers of a 42 inch screen was minimal for a viewing ratio between 3 and 4 (Sakamoto *et al.* 2008). The mental stress operationalized through physiological measures reached a maximum at 3H. Viewing distances between 2H and 3H provide the strongest feeling of involvement but people found viewing at 2H tiring, especially if the video included a lot of motion. Participants reported that their preferences depended on content and mood. A viewing distance of 4H was considered more relaxing by some. Some researchers, e.g. Poynton claim that for HDTV people will become familiar with and expect a viewing ratio of 3*H* (Poynton 2003). On mobile devices, people could easily enjoy HDTV at the design viewing distance of 3H on a screen of 13cm height (assuming a 40cm viewing distance).

Historically viewing distance was the only way to modify the angular size of the picture but it changed the angular resolution at the same time. New TV screens, computers and mobile devices are able to use only parts of the screen to display video and can therefore vary the size of the depicted content on the screen. Laptops and DVD players allow for native, full-screen or zoomed depictions of DVD content.

2.6.8 Angular resolution and angular size

In summary, displays provide a general raster of pixels of a certain density - also referred to as addressability - to depict content. The resolution of TV content is lower than the nominal addressability of the screen due to the camera line raster and the interlaced capturing process. For a given combination of picture height and resolution increasing the viewing distance has two opposing effects on perceived picture quality. The negative effect on the perceived quality is due to the picture angle becoming smaller in the eye of the observer. At the same time, however, the angular resolution of the pictures in terms of TV lines or ppd increases and thus improves the perceived quality, as long as the observers are not at their visual acuity discrimination threshold. A combination of viewing ratio and resolution should therefore define the parametrical space sufficiently while considering both the minimal viewing distance (15cm) and the ordinary acuity limit of about 60ppd. Despite the inter-dependency of viewing distances, image sizes and resolution much research has addressed them independently. What is currently not known is which image resolutions provide the best support for the different screen sizes of mobile TV devices. Although most of the results of the studies presented below did only take some parameters of angular resolution, PVD and viewing ratio into account I list the matching values for the sake of convenience. For television content I assume that the resolution of captured content has to be adjusted by a factor 0.7 for interlacing and inter-line twitter. I can assume that content that is scaled down by a factor higher than 0.7 will yield nominal resolution – for example standard TV content that is converted to a resolution of 320x240 can be assumed to achieve that resolution given the allocated encoding bitrate is large enough.

Jesty (1958) showed that people make use of increased sharpness of a TV picture by sitting closer to the screen. Their PVD was smaller for focussed slides than for slightly defocused slides. He found evidence for what he called an optimal viewing distance. When faced with the decision of placing a chair to view projected still pictures with a fixed size, observers chose their viewing distance in a way that depended only on the resolution of the picture. The quotient of picture height and optimal viewing distance was constant for a given resolution. Ribchester pointed out that Jesty's results could be entirely explained by conditioning to typical viewing ratios for 405 line pictures in their homes. In an experiment he let participants choose their PVDs for different size TV sets that were turned off. The results mirrored those of Jesty's studies in terms of PVDs but did not depend on resolution (Ribchester 1958) since no picture was ever shown.

Findings by Westerink & Roufs (1989) confirmed the existence of an optimal viewing distance that maximizes subjective image quality as posited by Jesty for moving images. But their results showed that at a constant viewing distance subjective picture quality was influenced independently by both the resolution of the picture and their size. This optimal viewing distance, however, was not evaluated as the preferred distance at which participants wanted to watch the projected pictures in a darkened room but was based instead on the computed maximum of the picture quality ratings provided by the participants at viewing ratios between 3H and 11.3H. The maximum quality was attained when the resolution equalled 16 cpd (32ppd) independent of the picture height. This indicates that the gains in perceived visual quality from achieving a higher visual resolution beyond 16cpd are not big enough to compensate for the reduction in picture angle. Visual quality higher than 16cpd only led to higher perceived quality if the size was increased, too. The studies by both Jesty and Westerink & Roufs were based on pictures that did not contain a line or raster structure.

For the reported values by Nathan *et al.* in 1985 the viewing distance (2.25m) of younger viewers translated into a visual angle (VA) of 12.3° of the picture with an angular resolution (AR) of 29ppd (adults: VA=8.5°, 43ppd). These values were based on NTSC screens with 575 lines and assume a Kell factor of 0.7.

Lund (1993) found no evidence that a person's preferred viewing distance depends on the resolution of the video material presented on TV sets or through projectors. The preferred viewing distances did not depend much on the resulting angular resolution of the picture. Therefore, he concluded that PVD is not a good indicator of visual quality. Correspondingly, line visibility did not have a large effect on PVD since it varied hugely across the preferred viewing ratios observed in his experiments. He looked at minimal viewing distances but due to his focus on VR he did not consider angular resolution as a limiting factor. However, in two of his studies in darkened rooms the minimal angular resolution observed must have been around 12ppd according to my calculations (see Figure 13). One was based on the minimal viewing distance at which people were willing to watch NTSC content, the other the PVD of half-NTSC resolution content. Both could have been limited by the effect of low angular resolution and line visibility.

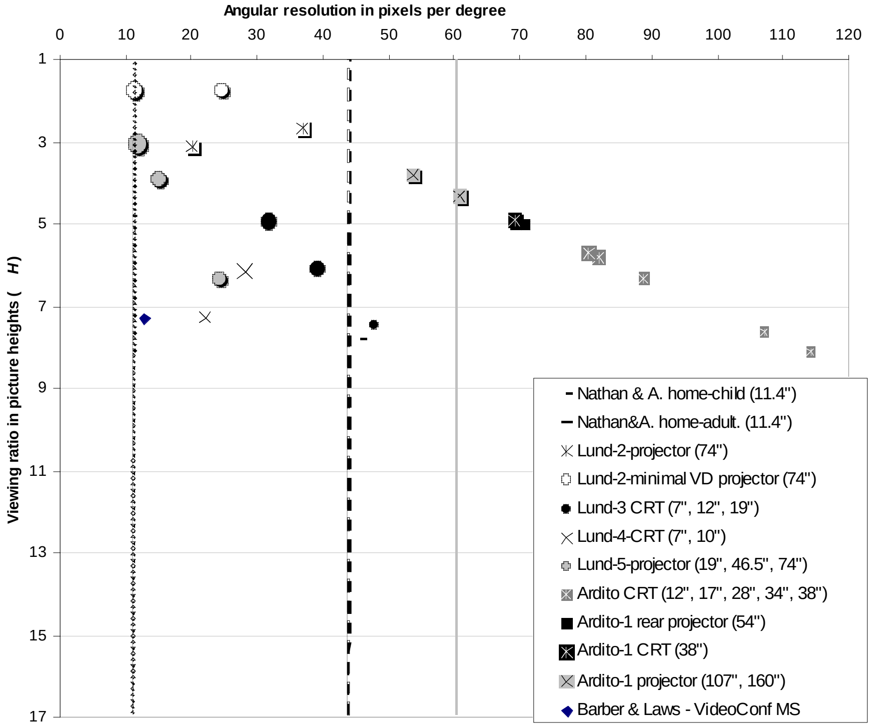


Figure 13: Preferences and limits of angular size and resolution combinations. Within the series of a study larger symbols denote larger screens.

In a number of studies on SD- and HDTV content Ardito, Gunetti and Visca (1996) found that if the contrast ratio of the content was not high enough, i.e. 109 or smaller, when participants watched content at the design viewing distance of 3*H* the added benefit of the higher resolution of HD content was lost. Ardito found that by reducing the ambient lighting, people sat closer to the screen. When watching HDTV content on a 38-inch diagonal screen in a completely dark room the average preferred viewing ratio was 3.8 compared to 6.3 when viewing the same footage in a lit room. At the participants' PVD of 5.2*H* the increased definition of the HD display was above human discrimination. Barber & Laws (1994) suggested that for a speech-related task of a health care video communication application a resolution of 128x128 should be avoided for pictures subtending visual angle of more than 10 degrees (VR≤7.3). This equates to a lower bound of angular resolution of roughly 13ppd - close to the values I derived from Lund's results but in a very different context (see vertical dotted grey line in Figure 13).

According to Birkmaier (2000) approximately 22cpd (44ppd) is perceived as a sharp image (see vertical dashed black line in Figure 13). This value is achieved when a typical TV display with 576 vertical lines (considering a combined Kell and interlace factor of 0.7) and a screen height of 50cm is viewed from a distance of three meters. Neuman (1988) found that at 7H 89% of naïve participants preferred standard NTSC (44ppd) over HDTV (89ppd) depictions. Since the former is close to w hat people are used to in the home this preference might be due to conditioning.

Figure 13 presents the collated information from the results on PVDs by Barber & Laws, Nathan & Anderson, Lund, Ardito and Ardito *et al.* along the viewing ratio in picture heights and the resulting angular resolution. The range of angular resolution at which people are willing to watch video content is large - from their ordinary acuity limits of 60ppd and above down to roughly 12ppd. Screen size and ambient lighting seem to influence the decision on viewing distance more than resolution. The preferred angular sizes and resolutions of depicted objects in passive viewing contexts associated with pleasure and entertainment on mobile devices have not been the topic of research.

2.6.9 Audio-visual interactions

As a by-product of a study on HDTV viewing experience, Neuman *et al.* (1991) discovered that the perceived video quality was improved by good audio. However, this was only the case for one of the three used content types. Similarly, a study by Beerends *et al.* (1999), using a 29cm monitor, found that the rating of video quality was slightly higher when accompanied by CD quality audio than when accompanied by no audio. In contrast, in the same study assessors judged two low video quality levels worse when they were presented with audio, than when the video was not accompanied by audio. The effect, however, was small. The violation of additivity of media quality will be discussed in more detail in Sec. 3.8 and Sec. 3.8.1.

Reeves *et al.* (1993) carried out an extensive study on people's responses towards audio-visual content. They tested attention (operationalized by reaction times to a secondary task), memory (correct responses to questions about auditory and visual content) and evaluated their experience, e.g. enjoyment, involvement and effort (through pen and paper based Likert scales). The study included forty participants in four between-subjects conditions (audio and video size) and four within-subjects conditions (high and low audio and video fidelity). Low audio fidelity led people to pay significantly more attention and better remember both audio and video content. There was a trend to rate picture quality higher when presented with low audio quality. Furthermore, with low audio quality the participants' ratings indicated being more part of the action, the scenes more exciting, the action faster and a stronger reaction towards the content. High audio quality was more realistic but required more effort according to the subjective questionnaire ratings. The study found many interactions of audio quality with participants' gender e.g. on attention, auditory memory, whether the depiction appeared realistic and recommended further studies on the matter. This is further justified by the large number of conditions and the comparatively small number of participants, which calls into question the validity of the statistical results.

2.7 Summary and discussion

From the described results it is clear that there are a large number of factors that influence people's perception and preferences when it comes to the optimal way to watch video content on TV screens and possibly mobile devices. Much of the research presented in this chapter is potentially limited in its external validity. For example, still pictures were used in many studies of video quality (e.g. Westerink & Roufs and Jesty) and viewing preferences (e.g. Tanton) for pragmatic reasons– displays with the required range of resolution were not available. The applicability of the results for mobile devices is therefore not guaranteed if they are not based on successively shown slides as in cinemas but on rasterized and interlaced camera sensors. Gouriet (1958) questioned the external validity of Jesty's research on PVDs because the used still pictures did not exhibit such visual artefacts as line break-up and stroboscopic effects that can occur in rasterised video. Ardito *et al.* (1994) found that the preferred viewing ratios for still pictures and moving pictures are different. The preferred viewing ratio was slightly higher for a moving picture in comparison to a still picture. In their lab trials people preferred slightly smaller moving images with a higher angular resolution and the pictures' contrast determined the degree to which they preferred smaller images.

But in summary the presented work suggests that:

1. for TV in the home viewing distance can be considered fixed. People currently do not adjust their viewing distance even with the arrival of larger screens in the living room;
2. in the lab the PVD depends mainly on the size of the TV screen;
3. viewing distance might have an effect on perceived video quality and should therefore be chosen according to realistic settings for mobile multimedia consumption;
4. in the lab TV viewing ratios smaller than 5H are only chosen in darkened rooms;
5. on large screens in darkened rooms people do not mind watching content at angular resolutions much smaller than 60ppd ordinary acuity limit and the 44ppd sharp picture mark - down to 12ppd;
6. 44ppd constitutes a sharp image but increases in angular resolution beyond 32ppd at the expense of angular size had an adverse effect on the perceived picture quality (both obtained at viewing distances of 2.9 and 5.4 meters, which are not representative for mobile devices);
7. the use of still pictures instead of moving pictures - although unavoidable in many studies due to technical limitations - imposes several limitations on the external validity of the results in terms of resolution, and addressability;
8. the influence of line or pixel visibility is not sufficiently understood for PVD and video quality. The size of the pixels might interact with people's preferences on their preferred angular resolution. Many previous studies did not used raster or line based displays – different resolutions were achieved by defocusing the lens of a projector presenting the material;
9. five fps is the lower limit for frame rate but it is not preferred for entertainment consumption. People do not seem to object to frame rates between 10 and 15fps on monitors but 12fps might be a possible cut-off for mobile device use.
10. video encoding bitrates by themselves and in conjunction with audio quality, text quality and frame rates that jointly deliver a mobile TV experiences are not understood;
11. if categorical scales are used in the assessment measures need to be taken to correct for uneven interval sizes (see Chapter 3);

As outlined in the above list, some conditions under which the previous research results were obtained might limit their application to the domain of mobile devices. The used viewing ratios (e.g. 6*H*) yielded relatively large angular sizes in comparison to what is possible in mobile contexts (e.g. 10*H* and higher) Almost none of the studies did include a way to provide qualitative feedback in the various studies that looked into the effects of screen size, resolution and video quality.

The results of QoE assessment should provide us with results that can be used to maximise the QoE for a given parameter set. In order to carry out my research on QoE in mobile multimedia applications order I need to know how to best measure visual quality. The next chapter reviews the scope of QoE and the techniques to assess video quality along with their limitations.

Chapter 3

Thoughts without content are empty,
intuitions without concepts are blind.
- I. Kant

Methods - measuring Quality of Experience

This chapter explains how my research fits into the greater research context. What am I exploring and how am I trying to measure it? In a competitive market place companies achieve their survival by producing products or providing services that customers are willing to pay for. Consequently, the process of designing products and services that will be adopted by customer has received significant attention and resulted in the formalization of e.g. the human-centred design process. Different disciplines have identified the need to better understand the value people attribute to products and services. The concept of experience promises to capture people's perceived value for network providers (Quality of Experience), human-computer interaction research (User Experience, UX) and industry (Customer Experience). All strive for better designs of products and services. Measurements of QoE should result in models and give rise to theories that can compare different designs of products and services and predict to what degree people will find them valuable in a given context to address their needs. Models of QoE should be able to predict people's choices. So far, however, video quality has been the predominant operationalization of QoE of services displaying moving images. This has resulted in a sizable body of research capable of predicting video quality ratings from trained observers under lab conditions. Undoubtedly, video quality contributes to the experience of multimedia services. But video quality has failed to explain many choices and trade-offs that people make when consuming video content, some of which I explained in the previous chapter. I will review approaches that help explain people's expectations and choices in technology and focus on measuring the contribution of video to the acceptability of the visual experience of a multimedia service on mobile devices. The different media assessment approaches are described along with their respective restrictions and problems that might render them inappropriate for certain applications. Based on the review of assessment methods I will justify the method that I chose for my studies.

3.1 Motivation

My research on Quality of Experience in mobile multimedia services intersects several disciplines: Human computer interaction (HCI), focussing on discretionary use of fast evolving computer systems and services, and the long-term standardization driven work of the (tele-)communication and motion picture engineering communities. The disciplines share the desire for economically efficient solutions to different degrees and have operationalized different core phenomena. For example, a predominant measure in the field of multimedia research is that of video quality, but as shown in Chapter 2 it falls short of explaining people's preferences under various conditions. I will briefly describe how these disciplines have all found it necessary to extend their focus to include a notion of experience.

Experience is a subjective and comprehensive concept that might seem too elusive and vague to support an inquiry aimed at attributing value to designs. Various researchers have suggested a number of high level concepts that need to be included to fulfil the ambitious scope of UX and QoE such as aesthetics, hedonism, emotions, context and social factors. However, clear suggestions for operationalizations of these concepts and their overall integration into QoE and the feasibility of including this into product and service design are missing. Research has trialled some initial operationalizations of concepts such as *sense of reality*, *larger than life*, excitement and enjoyment, described in Chapter 2, to better describe and quantify experiences and explain people's choices of technologically mediated experiences. But so far this effort has been fragmented by domains and somewhat random angles of inquiry. So far, QoE is but a goal (though not properly defined) that widens the scope of traditional multimedia research beyond objective and subjective audio and video quality. Although audio-visual quality should constitute an important factor to the experience of, e.g. IMAX and 3D cinema it does not address the difference in experience of these to standard cinema or TV.

Mobile communication and entertainment content address various needs people have, for example, to stay in touch and relax. Previous concepts for service requirement specification, such as network-centric QoS, have not adequately addressed the perspective of the targeted users of these services. They did not capture people's needs as common in human-centred design and the value that they attribute to a service within a given context. Service designers lack substantial knowledge on how people want to use mobile multimedia services and which parameters shape and define the experience that they would value the most. An understanding of the value of these experiences will help predicting the choices that customers exercise in light of a competitive market.

Designers and engineers aiming at QoE need concrete help in achieving these more ambitious goals specifically in identifying and measuring its defining factors. Science aims at describing empirical phenomena and at developing rules and theories that are fit to explain and ultimately predict them. Measuring the phenomena allows for the application of mathematical concepts to be applied to empirical data. QoE begs the question which phenomena need to be measured and how. My work focuses on the visual experience that forms part of QoE and how to measure it.

3.2 Human-centred design

The roots of human-centred design (HCD) lie in the Scandinavian school of cooperative design of information technology (Greenbaum & Kyng 1991). HCD aims at achieving designs that result in products or services that meet user needs and requirements through an iterative approach. The international standardization organization (ISO) standardized HCD and its procedures in ISO 13407 (1999). By identifying a need for HCD the designers of a new technology start an iterative process. An understanding and specification of the context of use informs the specification of user and organisational requirements. This leads to the production to a design - typically a prototype - that undergoes evaluation against the specified requirements. If the design meets the specified requirements this design cycle ends if not another iteration starts. An additional benefit of HCD is that when user needs are included in the evaluation of designs against requirements this potentially spawns new ideas for future services and research. As I will show in Chapter 6, Chapter 9, Chapter 10 and Chapter 11 participant feedback was instrumental in driving forward my research.

A user's activity does not occur in a vacuum but within a distinct context. The starting point of human-centred design according to ISO 13407 consists of the identification of the intended use contexts. A use context can be described by the user characteristics, the activity and the environment. The user characteristics include knowledge, skill, experience, education, training, physical attributes, habits, preferences and capabilities. The environment includes the technical, physical, ambient, legal, social and cultural environment. That is a long and potentially conclusive list of factors that designers can take into account - many of which will be dealt with by different actors in the inception of a technology. Dey *et al.* (2001) define context more application specific as the information that are relevant to the interaction between user and application. In their opinion places, people and things are the relevant entities and identity, location, status and time are the characteristics of context information.

In the case of multimedia services with passive consumption of content many methods for interactive tasks or usability are not suited – see Chorianopoulos (2004) for a discussion. For example, Drucker *et al.* (2004) showed that people preferred the user interface of a TV service with the worst efficiency and effectiveness. The preference was based on the fun the users had and how relaxing it felt to use the service. Wilson & Sasse (1999) had included user cost such as stress into the evaluation of multimedia services through physiological measures.

Mobile multimedia services differ from those in home or other stationary setups in many ways. The most obvious difference lies in the portability of the device with its resulting limitations in power supply, physical dimensions and the interference with other contextual factors both in the visual and the auditory domain, e.g. reflection and noise. Portability calls for wireless networks, and while using these networks the user faces limitations in terms of geographical positions and velocities in 3D space. Networks are the means to deliver content through a variety of services. However, services that include the possible presentation of continuous media possibly in real time to a range of customers still represent a technical challenge for the networks and often result in suboptimal presentations. Mobile devices change the landscape of media consumption since the devices tend to be personal and can change location with the owner. The prevalent use of TV was that of a shared screen in a shared space. I will discuss this in more detail in Sec. 4.3. Corporate research on mobile context follows a pragmatic approach that considers the technical, physical and social context in mobile contexts of use (Väänänen-Vainio-Mattila & Ruuska 2000). Their definitions concentrate on the limitations of the mobile environment that might impair the users' QoE.

Criticism of human-centred design in the field of computer science has targeted the fact that it does not match well with current software engineering approaches (Gulliksen *et al.* 2003). Although many different software design approaches exist, the most common one still follows an initial requirements analysis, which is then condensed into some model of the application and data structures. Norman has criticized HCD for being overly concentrating on people rather than the activities that people wish to perform through technology (Norman 2005). He claims that people adapt to tools in order to perform activities and therefore the activity should be at the centre of the requirements analysis.

3.3 Technology adoption

The adoption of new technologies depends not only on the experience they offer or the needs they address but on people's attitude towards new technology in general. Rogers (1995) categorised these different adopters into five categories: innovators, early adopters, early majority, late majority and laggards. He showed that people evaluated innovations' according to their advantages over current practices, whether the technology is compatible with their other activities. The perceived complexity of an innovation, whether it can be tried before committing to a purchase and whether other people will be able to observe them using the innovation affect customer decisions further. In section 3.6.1 I will describe how the attitude towards adoption affects people's ratings of video quality.

Behavioural sciences have been trying to understand and predict human behaviour in general - not only with respect to their use of tools and products. Ajzen & Fishbein's (1980) based their Theory of Reasoned Action (TRA) on the realization that people's *behaviour towards* a given *target* could not be satisfactorily explained through people's *attitudes* towards that *target*. People's *attitudes* and *subjective norms* - towards the *behaviour involving the target* delivered better predictions. The two most salient aspects of TRA – the focus on contextual use and social considerations - are of particular interest since they come up in many definitions of UX or QoE. In their model subjective norms describe the social pressure people perceive - the beliefs about what other people that matter to them might think about them performing or refraining from the behaviour involving the target.

Researchers have tried to explain why people adopt certain technologies and which factors determine users' acceptance of new technologies in a work context. Davis' (1989) Technology Acceptance Model (TAM) and its successors (Venkatesh *et al.* 2003) - explained employees' acceptance of information systems (IS) in the work place through important extrinsic factors such as *perceived usefulness* (PU) and *perceived ease-of-use* (PEOU). Davis defined perceived ease of use as *"the degree to which a person believes that using a particular system would be free from effort"* and perceived usefulness as *"the degree to which a person believes that using a particular system would enhance his or her job performance"*. The dependent variables that TAM considered were usually frequency, duration and variety of system functions used (Benbasat & Barki 2007). TAM showed that a range of factors affected the acceptance of a novel technology but it did not provide any guidance as to how these affect adoption or how PEOU or PU should be operationalized.

3.4 Roads to experience

The concept of experience and interest in its measurement is by no means new. Bentham's utilitarian theory on happiness targeted hedonic qualities that can be likened easily to the notions of *experience.* Edgeworth (1881) even fantasized about "*a psychophysical machine, continually registering the height of pleasure experienced by an individual*". In economics the concept of utility has a similar scope. A large body of research exists on people's remembered and experienced utility of an episode, see (Kahneman *et al.* 1997) for an overview. One of the major problems identified in this body of work in economics is that experienced utility of an episode as reported in real time (instant utility) and aggregated over time differs from retrospective evaluations (remembered utility). Duration neglect has been identified as the main problem. Although people agree on the fact that duration is an important factor to quantify utility it does not factor into remembered utility. As I will describe later this phenomenon exists in video quality research, too. Using experience as a yardstick for perceived value and as a predictor for human choice has been implicitly promoted and pursued by network providers, the human-computer interaction research community and industrial outfits.

3.4.1 Industry - from services to experiences

Throughout history people have employed tools to extend the capabilities of their bodies in order to tackle problems and satisfy needs, which proved impossible or required too much effort otherwise. Appropriation of physical tools to tasks outside their original scope was limited due to their physical properties and contextual constraints. Consequently, people and companies have invented a range of products and services to address their needs in given contexts. However, many if not most inventions and innovations are not adopted. Companies, whose existence depends on satisfying customer demand through their products and services, drive innovation but need to limit its cost and risk of failure. Explanations that link specific factors of an innovation to explain its adoption by a target group are difficult but research efforts have addressed parts of the problem. Pine and Gilmore (1999) posited a shift from a service economy to an experience economy as enhanced experiences are deemed harder to provide than services.

Company white papers e.g. Polycom (O'Neill 2002), (Empririx 2004), Nokia (2004) and Intel (Beauregard *et al.* 2007), define QoE typically as a comprehensive concept, encompassing all the stages in which a user might discover, purchase, use, maintain and dispose of a service or product. During the use stage some definitions mention usability, fun and pleasure (Beauregard *et al.* 2007) as important criteria but do not suggest how to measure these. Based on a task that customers might want to solve with a product and what matters in that context Aldrich *et al.* (2000) propose a definition of QoE as:

> *"What a customer experiences and values to complete his tasks quickly and with confidence."*

Companies are interested in benchmarking their products and services to be able to measure and compare them (Aldrich *et al.* 2000). Nokia (2004) focuses on the key performance indicators: reliability, availability, scalability, speed, accuracy and efficiency and their contribution to QoE. They define QoE as a concept not a metric but quantify it through labels e*xcellent, very good, good, fair and poor*. The value chain that influences these factors includes the service and network providers including network infrastructure providers and system integrators, user device and application software, and the provided content.

3.4.2 HCI - from interaction to user experience

The proliferation of computers, especially since the arrival of personal computing in the 1980s, has enabled people to tackle an ever-increasing number of problem areas such as office work, design and media production in a more efficient manner. Instead of having many tools each specifically designed to be used for a certain task only, computer hardware has basically provided the Universal Turing Machine, which, given the right software and input and output devices, can be used to address any information processing problem. The widespread use of computers in information and communication technology as well as in many other traditionally mechanical devices bears witness to the flexibility of computers as tools and the paradigm of the universal machine. The adaptations of an increasing number of activities with varying degree of success to this available platform prompted research to improve human-computer interactions to perform tasks of a certain problem class in digital environments. Hewett *et al.* defined the scope of HCI as "*concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them*" (Hewett *et al.* 1992). It considers the mechanisms inside the systems as well as the concerns that emanate from the people using them. Especially, computer science, psychology and ergonomics (Nickerson & Landauer 1997) but also sociology and cognitive science have informed human-computer interaction (HCI) research.

In the early days of computing nondiscretionary use of computing machinery was the norm and its division of labour resulted in three distinct activities: operation and data entry, management and programming (Grudin 2005). *Human Factors & Ergonomics* research focuses on operation and data entry and its scope can be summarized through Shackle's (1990) framing of *usability*, that included *effectiveness, flexibility, learnability* and *satisfactory use*. Human-computer interaction from a managerial perspective sought to answer how to design and introduce *Information Systems* (IS) that employees would accept and use. The computer human interaction (CHI) research community evolved once discretionary use of computers became possible and focussed more on users' "*initial experience"* (Grudin 2005) with computer systems, which unlike in IS they have to pay for. Research strands such as *ubiquitous computing* (Weiser 1991)*, pervasive computing* (Husemann 2001)*,* and *personal technologies* (Frohlich *et al.* 1997) have evolved that incorporate people's lives into technology design. *Ambient media* (Wisneski *et al.*, 1998), *information appliances* (Norman 1999), *invisible computing* (Weiser & Brown, 1996), and *the disappearing computer* (European Commission, 2000) address interaction with mainly discretionary computer systems' that people do not associate with regular desktop computer use.

A common criticism to the prevalent usability research was its focus on avoiding design mistakes and negative emotions according to usability criteria (Fulton Suri 2009) and not providing answers to why people adopted and used certain technologies. Since the mid-1990s HCI's mainly behavioural scope (Logan 1994) has been extended to explicitly include *user experience* (UX). Shneiderman (2003) pointed out that new technologies need to support relationships and activities that enrich people's experience in order to be successful. According to Stewart (2008) the International Standardization Organization (ISO) has considered issuing a definition of UX. Although still fuzzily defined UX in the field of HCI includes emotions, expectations, user needs and pleasure. Similarly, UX is included in *funology* – another recent branch in HCI research – that focuses on enjoyment and fun of the user (Blythe *et al.* 2003). Inspired by Maslow's *Hierarchy of Needs* Jordan (2000) proposed an extension of usability based on *Functionality, Usability and Pleasure.* Vyas & van der Veer (2006) classified four different kinds of experiences: sensorial, cognitive, emotional and practical. Hassenzahl distinguishes between do- and be-goals of user experience or in other words pragmatic or functional and hedonic quality.

McCarthy & Wright's (2004) framework for analysing UX includes four threads that make up the *braid of experience*: the compositional, sensual, emotional and spatio-temporal. To make sense of experience they suggest paying attention to anticipating (expectations), connecting (association and comparison with other services), interpreting (reflective monitoring and adaptation to ongoing experience), reflecting (value judgments), appropriating and recounting.

3.4.3 Networks - from Quality of Service to Quality of Experience

Mobile multimedia services can provide services such as mobile TV that people find entertaining and enjoy using. Mobile services require a mechanism that enables consumers to receive information from a sending entity wirelessly. Typically, this mechanism is provided through one or more networks. The use of interactive services and the just-in-time delivery and consumption of content across the network, pose a number of performance demands for networks. If the networks cannot achieve this performance, services might be impaired or rendered useless.

The adoption of QoE in networked multimedia research marks a shift from focussing on network performance and providing services to a more human-centred orientation. The shift is from "What services can we offer?" to "How do we design and provide services that result in optimal user experiences?" The latter challenge lies in designing and providing enhanced experiences for the end-user (Baker 2006) whereas the research on the former technology-centred provisioning has produced a number of ways to guarantee QoS.

Historically, there have been two distinct strands of networking – circuit-switched and packet-networks. The older circuit-switched approach creates a connection between two entities by providing a dedicated, albeit nowadays virtual, circuit from sender to receiver. If there are not sufficient resources available along the necessary path through the network to allow for the establishment of this line, users will be rejected – e.g. by not receiving a dial tone. Successful operators have to dimension their networks to minimise these disappointments while at the same time keeping over-provisioning within their commercially viable limits. Once established, however, a line guarantees a certain amount of information to be transferable with a certain QoS profile according to the technical specification of the system.

Packet-switched networks follow an approach resembling the postal system. The Internet is the most prominent example of a packet switched network. Packets are sent from sender and are routed through the network to the receiver based on address information and routing rules that adapt according to traffic and node availability. In its current inception the Internet provides a *best-effort* service, as packets might never arrive at the receiver. The fault tolerance, adaptability and dynamic reconfigurability of the Internet results in packets reaching the receiver with varying delays, and at times not at all. This is not much of a problem for *elastic* services, which can cope with varying and possibly lengthy delays due to data retransmission (Shenker 1995). But the delivery of continuous media is more sensitive to these impairments, which result in e.g. audio skips and freeze frames. From the beginning engineers and researchers have looked into providing Quality of Service (QoS) guarantees for the delivery of inelastic services through the Internet. So far there is no support for end to end QoS guarantees in the Internet (Hardman *et al.* 1995).

The CCITT defined QoS in E.800 (CCITT 1988) from a user point of view through their satisfaction with the performance of a service as:

> *"The collective effect of service performance which determines the degree of satisfaction of a user of the service."*

However, the more dominant operationalization of QoS was through measurable network performance parameters as defined in recommendation X.902 (ITU-T 1996) and not through the user's satisfaction:

> *"A set of quality requirements on the collective behaviour of one or more object in order to define the required performance criteria."*

Closing the gap between these two extremes the ITU has provided user-centric QoS categories that address user needs based on service response times and error tolerance as illustrated in Table 4.

Table 4: ITU model G.1010 for user-centric QoS categories

	Interactive (delay <<1 s)	*Responsive (delay ~2 s)*	*Timely (delay ~ 10 s)*	*Non-critical (delay >> 10 s)*
Error tolerant	Conversational voice and video	Voice/video messaging	Streaming audio and video	Fax
Error intolerant	Command/control (e.g. Telnet, interactive games)	Transactions (e.g. E-commerce, WWW browsing, Email access)	Messaging, Downloads (e.g. FTP, still image)	Background (e.g. Usenet)

The key parameters used by the ITU in (ITU-T 2001) to describe human requirements for different applications are:

- delay (which includes all aspects of transport and other parameters, e.g., bandwidth)
- delay variation (also referred to as jitter)
- information loss (which includes bit errors, packet loss, and also coding artefacts)
- degree of symmetry (one-way or two-way communication)
- miscellaneous requirements like adequate echo control, synchronisation between streams and packet loss concealment

Based on these one can define multimedia services such as Internet protocol TV (IPTV) SDTV and HDTV (*cf.* Table 5) in conjunction with the employed codecs.

Table 5: QoS parameters for different TV services

Service		*Bandwidth*	*Delay*	*Jitter*	*Packet loss*	
IP TV	One-way	512 kbps	< 10s	< 1 ms	< 1% PLR	Lip-synch: < 80ms
SD TV (MPEG2)	One-way	2-3 Mbps	< 10s	< 1 ms	< 1% PLR	Lip-synch: < 80ms
HD TV (MPEG4)	One-way	10-12 Mbps	< 10s	< 1 ms	N/A	Lip-synch: < 80ms

A set of QoS parameters as provided in Table 5 does not provide a good description of the service in terms of what users perceive and experience when using it. Too many choices both in terms of the overall service design and trade-offs between parameters are open for optimization and can result e.g. in very different perceived video quality. Consequently, although supported by the same set of QoS parameters people's experience of services can hugely differ according to user context, application, content, encoders, user perception and preferences.

This shortcoming has been identified and research on Quality of Experience (QoE) has been embraced to provide better predictions of people's overall preferences and to overcome the existing gap between mediated multimedia experience and QoS parameters and e.g. its resulting video quality. The ITU has adopted the term of Quality of Experience and the ITU-T study group 12 is studying and defining QoE in

Question 13/12. According to recommendation P.10/G.100 (ITU-T 2008) the ITU defines QoE in accordance with an earlier definition proposed by Ericsson (2006) as

> *"The overall acceptability of an application or service, as perceived subjectively by the end-user."*

Currently, the ITU has not defined how this *overall acceptability* is operationalized but the definition suggests that measurements will be obtained in the context of an end-user using an application or service. Siller & Woods (2003) made an attempt to map QoS to QoE. Their approach quantified QoE of the interface between a user and a system and included the effects as introduced by the network as:

> *"... the user's perceived experience of what is being presented by the application layer, where the application layer acts as a user interface front-end that presents the overall result of the individual Quality of Services".*

Their definition of QoE was inspired by IT companies' white papers (see Sec. 3.4.1) and similar to earlier approaches aiming at differentiating between QoS of application (QoS-A) and network (QoS-N or N-QoS) by e.g. (Campbell *et al.* 1993). Bauer & Patrick (2002) suggested an extension to the seven layer OSI (Open System Interconnection) reference model - often used to abstract from the complexity of distributed systems - to include *display*, *human performance* and *human needs layers* on the top-most *application layer*. The *display layer* (8) includes all notions about the input- output devices and the user interface. Layer 9 encapsulates all concerns about *human performance* on information processing such as perception, memory, cognitive effort, and motor skills. The uppermost layer encompasses long-term and general *human needs*, which are independent of technology. Bauer & Patrick argue that QoE should be used to express the requirements that stem from these three additional layers.

Network optimization, load balancing and other network design deployment and maintenance tasks have increased the need for automated measurements. Objective quality measures (see Sec. 3.9) provide approximations of how people might perceive video quality. Since these measures can easily be correlated with QoS profiles they provide an easy shortcut to assess performance of networks and services and have contributed to uncritical use of video quality as a proxy for QoE.

Of all three avenues to experience the one from the network is the least developed since both HCI and commercial services have historically had a stronger focus on people. My own contribution to QoE is geared to further the understanding of experience in the field of network provided services. My research should help predicting people's experiential preferences as the network engineering and multimedia research communities are looking for hands on guidance as to which parameters to focus on and what to trade-off when optimizing QoE.

3.4.4 Criticism of experience focus

Criticism of experience as a goal of design and scientific inquiry comes from two different angles: its feasibility and its value. The first is the philosophical perspective held by e.g. McCarthy & Wright (2004) and Vyas & van der Veer (2006), which posits that experience cannot be created or commoditised. People live with and adapt technologies to their needs. The experience occurs on the boundary of the technology and what the people bring to it when interacting with it. Cockton criticizes the focus on experience as too narrow and misguided. He shares the above philosophical view but his main criticism is that the focus on experience distracts from the more important targets for design - values and needs.

My research that focuses on experience is not meant to suggest that human experiences can be pre-fabricated and packaged. Experiences involving mediated events will always be individual but the technology that facilitates the mediation can be designed in ways that will be more enjoyable and have more value to people than others. Currently, the focus from a technological point of view stops too early without considering the user context. QoE does not mean that *one-size-fits-all*. Knowledge about experience has to be able to integrate individual preferences. Service designer will benefit from an understanding to design more valuable services by making better decisions with regard to salient experiential factors that can be pre-configured or made to be adjustable to match people's individual preferences. My results are meant to inform design *for* enjoyable experiences and not *of*.

3.5 Media quality assessment

Content providers and network operators are interested in providing and delivering economically viable multimedia services at acceptable quality levels. Thus, human experience has been both the motivation and the yardstick for technical reproduction of multimedia content. In order to be able to reliably map human perception - and ultimately experience - to technical parameters, quality assessment schemes are needed to act as translators between user perception and possible technical implementations of services. Users have different media quality requirements depending on the application and task that they seek to achieve with it and in which context. Quality assessment schemes should therefore consider application and task context to provide realistic results. User quality requirements for different applications and use contexts are not always easy to ascertain. Historically, psychophysical scaling and later multimedia quality assessment techniques that I will introduce in this section have dealt with the problem of mapping perceptions and multimedia qualities respectively to scales.

There are many different ways to operationalize the concept of quality with or without assessors. The problem of finding viable operational approaches to media quality is tightly coupled with measuring problems. How can we quantify quality, and how can we measure it? Media quality assessment methods are supposed to provide answers to how humans perceive different dimensions of audio or video content. Whereas quantitative approaches represent the *lingua franca* for quality assessment qualitative methods pose a popular means to identify the dimensions along which humans identify and qualify quality. For an overview on the human visual system and the impact on video quality assessment - see Winkler (1999).

In quantitative multimedia quality assessment, the research community is divided into two camps. The objective camp advocates computerized and therefore automated, low-cost techniques to obtain measures that quantify the fidelity/quality of content without the need for lengthy and costly subjective assessments. They are popular for calibration and optimisation of services and systems as they theoretically allow for automated system tuning and monitoring or on-the-fly determination of optimal settings for a given service on a per-clip and even per-user basis. Some objective measures like the Video Quality Metric (VQM) and the structural similarity index (SSIM) (see Sec. 3.9) employ human perceptual models to derive results that better match results assessors would provide. The validity of the fit with human quality perception and judgment, however, leaves many questions unanswered as described in Sec. 3.9. The subjective camp favours assessors to measure quality. The operational definition of quality is therefore very different, as it involves obtaining information from assessors being exposed to media and the most attention in this chapter will be devoted to it and its problems. Since the pioneering work of

Weber on just noticeable differences (JND) in sensations (Weber 1836), and Fechner's conception of a psycho-physical relation in the 1800s (Fechner 1860), the field of psychophysics has been dealing with the problem of measuring sensations. For the most part this research has been centred on "one-dimensional" sensations, e.g., loudness of a sound, luminance of a light source, or intensity of pain, but the method of discriminating between two stimuli has been successfully applied within various multimedia assessment techniques. Lawrence and Marks use the categories *sensory distance* and *sensory magnitude* to roughly classify the different psychophysical scaling methods and theories. The former refers to the classical or "old" psychophysics based on Fechnerian discrimination thresholds whereas the latter - new psychophysics - quantifies the magnitude of a sensation through a person's verbal response, e.g. using numerical scales (Marks & Algom 1998). The concepts of *threshold*, *method of limits*, *method of adjustment*, *method of constant stimuli*, and the *forced-choice procedure* are techniques developed and used in psychophysics where intensities of sensations are being compared and quantified.

Even though multimedia perception is not a one dimensional concept, many of the psychophysical techniques have been successfully applied in its assessment. Subjective quality assessment approaches differ in many ways. The general idea is to subject assessors to various quality levels and obtain judgments from them through mainly quantitative methods. During an assessment the assessors are being exposed to a number of media clips – often referred to as stimuli - and provide some form of response about the quality of the presented clip. The obtained ratings describe people's perception of the respective quality levels much more accurately than objective measures (see Sec. 3.9). But one has to take care in obtaining these ratings since assessors are far from perfect (*cf.*, Sec. 3.6 and e.g., Attneave (1962)) and for example provide ratings based on contextual factors they assume (Engeldrum 2001).

This section classifies subjective assessment approaches based on a set of dimensions:

1. What kind of rating is collected? If scales are used, what kind of measurement levels do they afford and which granularity does the scale have?
2. How long do the stimuli last?
3. How often are ratings collected?
4. When are ratings of a stimulus obtained from the assessors?
5. How much cognitive involvement does the response require from the assessors?
6. In what context are the ratings of obtained?
7. Are the ratings absolute, or do the assessors make comparative judgments about a stimulus and a provided an anchor? What comparisons can they make from the range of qualities present in the stimuli?
8. Can experts be assessors?

3.5.1 Scaling

I will limit my overview of scales to those that are most common in multimedia research. Scaling by distance is based on the idea of comparing two stimuli and deriving units of perception based on their perceptual differences. Just noticeable differences, paired comparisons and acceptability ratings can be grouped into this category and are described below. In scaling by magnitude approaches assessors are presented with various quality levels and provide - usually quantitative – ratings for these.

The type of scales used to gather the assessors' ratings determine to a large degree the kind of analysis that can be performed on the ratings. Ratings can be binary, or present a range of values on a categorical, discrete or continuous scale. Rohrmann (1978) found that 5–point scales are preferred by assessors. If the granularity of the scale is too fine, assessors will not make use of the fine-grained differences because they do not have confidence in their judgments about small differences. Furthermore, assessors tend to avoid the extreme ends of the scale making for a more condensed rating.

Paired comparison

The parallel or paired comparison is based on a psychophysical technique called *Thurstone's Paired Comparison Scaling* (TPCS) (Thurstone 1927). It takes advantage of the assessors' best abilities – comparative judgments. Assessors are asked to decide, which of two alternatives is better with respect to some given criteria. Comparative judgments are easier for assessors than absolute judgments. Presenting the two alternatives at the same time facilitates a judgment that is not based on a comparison of an ongoing stimulus with a remembered anchor stimulus. TPCS result in ordinal ranking data.

Depending on the modalities and the dimensions in which the two stimuli differ, it might not always be apparent why an assessor prefers one stimulus to another. In the case of video, the presentation of two stimuli at the same time is problematic because of the unity of the assessor's locus of attention and the uncertainty what features or regions he is using to arrive at his assessment. Follow-up questions and debriefing might be necessary to determine the criteria the assessor was relying on in order to come to a decision. However, this kind of disambiguation is rarely employed.

In the Double-Stimulus Continuous Quality-Scale method (DSCQS) a reference and an impaired stimulus are presented to assessors in parallel and the assessors provide ratings for both on individual scales. In DSCQS the assessors do not know, which video is the reference or the impaired video.

Just noticeable differences (JND)

Just noticeable differences are a concept that dates back to the research of Fechner in psychophysics done in the eighteen hundreds (Fechner, 1860). His *Method of Limits* was aimed at detecting thresholds in human perception in a single dimension. The intensity of a stimulus e.g. loudness was increased in discrete steps, until it was just detectible to an assessor. The assessor would indicate the detection of a difference in the stimulus by a binary YES/NO response. Typically, this procedure would also be reversed, i.e. the intensity of the stimulus would be decreased to find out whether the threshold remained the same.

Acceptability ratings

McCarthy *et al.* adopted the *method of limits* in the concept of acceptability. They asked assessors to state, i.e. call out, when they thought that the quality of a stimulus became unacceptable or acceptable for consumption (McCarthy *et al.* 2004a). Similar to JNDs assessors only needed to make a binary decision. Acceptability ratings can be obtained during stimuli because the overhead is low and the decision is fairly simple. In McCarthy *et al.* the assessors could switch back between these two opinions continuously as a stimulus of varying quality levels was being presented. Assessment can also be obtained after the stimulus. The method could be analogously used for other quality thresholds, e.g. 'good or better' or 'excellent'. In a similar way Apteker addressed minimum thresholds for acceptable QoS parameters with a label called *watchability*, which indicated a lower bound on video quality (Apteker *et al.* 1994) akin to *acceptability*.

The ITU Absolute Category Rating (ACR) or Mean Opinion Score (MOS) scale

The International Telecommunication Union (ITU) uses five category scale as shown in Table 6 with and without the numerical labels to rate quality or its impairments (ITU-T 2004) in Double Stimulus Impairment Scale (DSIS), Degradation Category Rating (DCR), and Absolute Category Rating (ACR) approaches (Winkler 2009). The labels are used for the assessors to provide ratings, which are then mapped to their respective scores to aggregate mean values. In some cases the numerical values are used for rating.

Table 6: ITU-R ACR Quality and Impairment scales

Quality	Score	Impairment	Score
Excellent	5	Imperceptible	5
Good	4	Perceptible but not annoying	4
Fair	3	Slightly annoying	3
Poor	2	Annoying	2
Bad	1	Very annoying	1

0-100

ITU-R Recommendation BT.500 defines a 0-100 scale (ITU-R 2004) that gets often used for DSCQS testing and Single Stimulus Continuous Quality Evaluations (SSCQE) (Winkler 2009).

CR-10

A scale, which claims to overcome some of the problems (discussed in Sec. 3.6.4) of MOS, is Borg's CR-10 scale with intervals that logarithmically decrease in size. CR-10 was designed to measure pain and exertion in physical training exercises. According to the author it has not only interval but also ratio properties (Borg 1982). However, the scale has seen no use in media quality assessment.

3.5.2 Stimulus length

To rate continuous media such as audio and video, assessors have to be exposed to the stimulus for a certain period of time. The stimulus duration varies for the different assessment approaches and has ramifications for the obtained results. Since humans are not very good at integrating judgments over long periods of time, research in media quality assessment uses mid-length stimuli of up to 30 sec or short stimuli that last less than 5 sec. The ITU for example recommends lengths of 10 seconds or less (ITU-T 2004). The Video Quality Experts Group (VQEG) uses a test set of twenty 8-second clips (VQEG 2000) to represent a range of difference types of motions, content and camera position. However, Aldridge *et al.* (1995) observed that a test sequence length of around 10 seconds was not long enough for assessments of video impairments as the range of impairments that are typically found in ATM video was not adequately captured. The use of short stimuli excludes possible long term effects and assumes quality requirements and the value of quality to be constant over time. For example, it presumes that users would be willing to watch a full length feature film in a low quality that would be acceptable to them for a short advertisement. It seems plausible that people would opt for a higher quality if they committed to long multimedia content and have therefore higher quality needs depending on the assumed duration.

Stimuli that are too long have proven difficult, too. Longer stimuli may be boring for the assessors. In both (Aldridge *et al.* 1995) and (Watson & Sasse 1997) "*distracted*" quality assessors were reported.

3.5.3 Time of response

Depending on the point in time when the assessor gives a rating in relation to the presentation of the stimulus, one can classify these methods into *intra-stimulus* and *post-hoc* approaches. In the former case, the assessor can provide his answer during the stimulus period whereas in the latter case the stimulus has ended. The intra-stimulus rating schemes make people conscious of the rating process with continuous rating during the whole stimulus period. Irrespective of the point at which the feedback rating is collected from assessors, it affects the feedback. A judgment in intra-stimulus schemes might be prompted by an incident or might have built up over time. However, the post-hoc approaches are even more subject to biases e.g. recency, forgiveness, primacy, and ordering effects. These problems will be discussed in more detail in Section 3.6. Watson (1996) reported cumulative effects on the quality assessment of an hour-long video conferencing such that it was impossible to identify what periods were responsible for the rating.

3.5.4 Response frequency

In intra-stimulus approaches, stimuli can be rated once per presentation or continuously. The mere act of rating introduces a cognitive overhead that may indirectly affect the result depending on the effort required. A scale that has been poorly designed or does not map easily onto the presented variations of the stimulus can affect the mood of assessors, and thus their ratings. The rating requires stepping out of the context of consuming video. The cost in terms of time, the amount of steps, and the complexity of a step involved can affect ratings.

3.5.5 Cognitive involvement

Providing ratings requires attention and the required cognitive effort can influence the judgement. De Ridder & Hamberg (1997) found that with elongated stimuli recommended methods such as DSCQS could not be employed, since the load on memory became too great.

Non-intrusive approaches obtain measures without conscious involvement or cognitive effort on the part of the assessors. These measures are typically objective in the sense that assessors have no means of deliberately altering the results. I will explain two examples for this in Sec. 3.7.

3.5.6 Anchoring

The range of qualities that are used during an assessment influences the ratings since it implies feasible qualities and influences people’s expectations and is often referred to as the rating context. Assessment methods that explicitly use reference stimuli can be classified into *parallel* and *sequential* comparisons, depending on the temporal relationship between the original and the reference. For example, some non-reference approaches (also called *absolute* ratings), establish explicit anchors during the training part of an assessment session. When anchors are used in conjunction with a rating method - e.g. MOS - assessors experience the baseline or a range of qualities according to which they are supposed to judge the following stimuli. Typically this provides good results in terms of a low variance in the responses as training can overcome people's inability to detect and recall differences accurately. The major question that arises however is what are we measuring? The resulting ratings do not necessarily represent their absolute perception or appreciation of quality. The assessors might not agree with the anchors. This represents a real problem because ultimately, the assessment is supposed to capture perceived quality as is, rather than shape that perception. As Torgerson pointed out - if you want to use any resulting scale in conjunction with naïve users you should base the construction of that scale on naïve users (Torgerson 1958). In short, anchoring may remove variance of subjective ratings but thereby also removes the assessors' expectations an important part of the concept of QoE. However, even if no anchoring is employed the range of qualities which assessors encounter can become an implicit anchor. For example, an increase in audio quality lead to a better perceptual quality than a decrease to the same level in (Bertram & Steinmetz 1997).

3.5.7 Assessors' background

Similar to anchoring the assessors' background influences subjective quality ratings. Some assessment approaches require untrained assessors who have not been exposed to the kind of stimuli whereas other approaches are expert-proof, i.e. the ratings of expert assessors (of the dimension being investigated) do not differ from untrained assessors. One problem that arises with experts, who have had more exposure to the investigated phenomena, might provide more conservative ratings than non-experts. Experts look for degradations in video that might not be apparent or important to novices or regular users. However, untrained assessors are being confronted with a situation in which they are supposed to make decisions about something they usually do not think about. They provide ratings that they can justify and possibly make them look smart. These rationalised decisions can be far from what they would do in a natural setting where they have different choices about the quality of a service e.g. accepting it because of non-existing alternatives, trying to increase or complain about the quality or not using the service anymore. This largely depends on the context. Bouch (2001) showed that assessors' knowledge and experience of networks, and the real-world task they perform with applications, determined their ratings.

3.5.8 Context

Visual perception can vary along many different dimensions, e.g., colour, intensity, contrast, sharpness, size. In his image quality circle framework Engeldrum categorises visual attributes like sharpness and graininess and other so-called "nesses" that are responsible for image quality. He stressed that integrative attributes like image quality are more context- or application-dependent than these "nesses" (Engeldrum 2001). As Mellers *et al.* (1996) pointed out: "*Preferences do not occur in a vacuum, they are always formed relative to a context"*. Assessors' preferences and their judgments occur in a context, which may be clearly defined or implied by an experimental setup or assessment approach. In real world situations, however, users may react differently to quality levels than in the lab since their perceived quality is grounded in context, in-situ requirements and relative to their real life expectations.

3.6 Caveats in subjective assessments

As opposed to the automated objective approaches, human perception and information processing used for subjective assessment exhibits the following characteristics (Hogarth 1980), all of which may distort the obtained results:

1. Selective perception – e.g. the locus of attention acts as a filter.
2. Sequential processing, i.e., they exhibit temporal effects with non-linear distortions.
3. Limited memory – especially with longer stimuli people have problems integrating perceptions over time (see also Sec. 3.5.2).
4. Limited 'computational' ability.
5. They are adaptive and conceptualised with a dependence on task characteristics.

More concretely, a number of traits have shown to influence assessors' rating behaviour.

3.6.1 Assessors' expectations and interest

Assessors might not have direct experience but nevertheless expectations. Zeithamel *et al.* (1990) stressed that expected service quality affects its perceived service quality – it will be perceived as positive when expectations are exceeded, satisfactory when expectations are met, and negative when expectations are not met. This means that expectations of service performance may be tied to its pricing and the promises that the marketing of the service implies. Perceived quality therefore is not a constant. Perceived quality evolves as technology advances. Meeting expectations is subject to treadmill effects. Human adaptation to the status quo has been labelled the treadmill effect observed in the area of human satisfaction with e.g. quality of life (Kahneman 2003). As quality improves so do our expectations. Although human perception can be considered constant measuring the perceived quality will yield some reference to expected quality and therefore vary. Perceived quality has to be considered in relation to comparable available services and may rely on pricing (see next section).

The expected performance and experiences provided by new technologies are shaped to a large degree by the media and marketing efforts. This phenomenon has been formalised in Gartner's hype cycle model (Linden & Fenn 2003). Their studies indicated that the introduction of a new technology that gets adopted follows five distinct phases. The first is triggered by a technological breakthrough, product launch or other event that receives enough publicity and interest. In the second phase *'a frenzy of publicity typically generates over-enthusiasm and unrealistic expectations'*. A failure to meet the *inflated expectations* leads to a negative hype, which makes the technology quickly become unfashionable. The media usually abandon the topic, which enters the *Trough of Disillusionment*. Although the media may have stopped or reduced coverage of the technology, some businesses continue through this phase, improve the technology and try to understand the real benefits and practical application of the technology leading to a *Slope of Enlightenment.* As technology reaches a *Plateau of Productivity* its benefits become widely demonstrated and accepted by a large community. The technology becomes increasingly stable and evolves in follow-up generations. The height of the plateau varies according to whether the technology is broadly adopted or benefits only a niche community. However, the hype cycle model does not propose any operationalization of these concepts, which makes its application to QoE difficult.

Even if assessors are naïve with respect to a new technology they are not neutral towards the content presented during video quality assessments. They might be interested in the type of content or bored by it. Kortum & Sullivan (2004) showed that video quality ratings of different content types depended on the assessors liking of the content. People that were interested in a certain content type provided higher ratings than assessors that were neutral or disliked the content. To achieve greater validity of the results services, content and their qualities should be assessed with people that are interested in them. A recent study showed that innovators and early adopters - of Rogers' diffusion theory - have lower requirements in terms of video quality (Jumisko-Pyykkö & Häkkinen 2006).

Similarly, the assessors' cognitive style can influence their ratings. Cognitive styles describe human differences in approaching, e.g., the assessment of a picture based on the details *vs.* the whole. When these cognitive styles cannot be controlled for results will contain ratings from differing perspectives (Ghinea & Chen 2004). This suggest that assessment should be performed in a clear context including a task or service and that ratings alone may lead to an overly simplistic interpretations of the results.

3.6.2 Pricing

In most cases users have to pay for services. Bouch *&* Sasse (1999) showed that users' tolerance for QoS and their attitude to controlling payments thereof is governed by their level of confidence that the performance of the QoS parameters reflects value for money. Even if no pricing information is included in the assessment assessors make implicit assumptions about pricing based on their expectations (Bouch 2001). In cases in which users have no control over their payments for the obtained QoS this is governed by their degree of *Peace of Mind;* users accepted a lower quality if it was guaranteed through an agent and they did not have to get involved.

Möller (2000) found a linear correlation between MOS ratings and the willingness to pay for these. But he considered this result unrealistic and dismissed it as an artefact of the laboratory setup – deemed unfit to measure the interaction of QoS and pricing in the real world. No systematic research has been conducted to see how the pricing of a given video quality affects its perceived visual quality.

Quality requirements are tied to economic constraints. For example, Drucker defines quality as a customer's perceived value for money with respect to the characteristics of a commodity (Drucker 1999). For him *'Quality in a product or service is not what the supplier puts in. It is what the customer gets out and is willing to pay for. [...] Customers pay only for what is of use to them and gives them value. Nothing else constitutes quality'*. However, media content and their quality requirements have to date hardly been studied in relation to pricing largely due to the absence of adequate assessment techniques. Monetary incentives or payment schemes have proven hard to employ in lab experiments where people are given money at the start of the experiment and can spend this on higher quality levels - see for example (Hands *et al.* 2007). The question "What are people willing to pay for a certain level of quality?" remains unanswered and is usually left to the dynamics of the market. In particular, it is unclear what gains in value users perceive from higher media qualities. Service and network providers along with application designers not only need to know optimal parameters and the minimum quality required in a given context, but also the maximum point beyond which users see no added value in increased quality. On the opposite end, consumers might accept lower quality when coupled with lower cost (Podolsky *et al.* 1998).

To date the influence of pricing is hardly understood and therefore current assessment methods do not consider pricing.

3.6.3 Temporal Effects

Ordering effects can influence ratings when they capture impressions over time and of varying quality levels. A *primacy* effect is defined as "*when the message presented first exerts a disproportionate impact on an individual's opinion*" (Crano 1977). A *recency* effect is defined as *"when the later message predominates*" (Crano 1977). Hands & Avons (2001) showed that retrospective quality ratings were poorer when the worst-quality video occurs at the end compared to the beginning of a 30 second video clip. Correspondingly, assessors *forgive* impaired video when it is followed by a substantial period of unimpaired video (Seferidis *et al.* 1992). Other ordering effects are due to the fact that stimuli might be judged relative to their predecessor as noted by Sporer (1996).

Fredrickson & Kahneman (1993) showed that for subjective experiences *peak-end* values were good predictors of assessors ratings. In their case the highest level of pain experienced during a medical procedure and the level of pain at the end of it. Duration of pain did not factor into the post-hoc ratings but. Hands & Avons (2001) found evidence that the concept of peak-end values applied in video quality assessment, too. When assessors continuously provided intra-stimulus ratings of picture quality a recency effect was not present. Durations of impairments in video quality had little impact on post-hoc quality ratings. In the continuous rating process the video quality ratings were best predicted using the intensity of the peak impairment. Hands showed that duration neglect was present when assessors judged the quality of impaired video sequences. Duration neglect represents a serious problem as the aggregate of continuous ratings, differ from the remembered quality.

Assessors rated larger steps in the change of quality worse than smaller steps to the same level. Users prefered constant lower quality over an on average better quality with variations (Bouch & Sasse 2001).

3.6.4 Field Context

Evaluation in the field is time-consuming and difficult to carry out because experimental control in terms of interruptions, movement, lighting and sound conditions is hard to achieve (Tamminen *et al.* 2004). Especially for video quality assessment and their comparison the ITU guidelines suggest control of e.g. lighting conditions (ITU-R 2004) that is not feasible in the field. However, results that are obtained in more realistic settings will have greater predictive validity, as they are closer to the real experience in which the value of a service will appreciated (Sasse & Knoche 2006). The only other study – apart from study 3 presented in Chapter 8 - that compared video quality assessments in the field and the lab was by Jumisko-Pykköö & Hannuksela (2008). Participants rated the acceptability of video quality higher in the field than in the laboratory for all four tested error ratios (from 1.7% to 20.6%). Interestingly, the video quality with 1.7% error ratio received lower satisfaction (MOS) ratings in the field (6.5) than in the lab (7.5) but the participants' acceptability ratings for the same error ratio was higher for the field (89%) than in the lab (82%). This provides evidence that the measure of acceptability is not solely based on visual quality.

3.7 Other assessment approaches

There are other approaches that have been used to find out how media quality affects people, their attention and their ability to perform tasks. I will briefly discuss these and their potential contribution to measuring QoE.

3.7.1 Task performance

Task performance measures (TPM) are commonly used in *human factors* research on nondiscretionary technology use and capture the impact of media quality on people's ability to perform and complete tasks. Information transfer as measured in Ghinea & Thomas' model of QoP (see Sec. 3.8) is an example of a perceptual quality measure that is based on the performance of subjects. Rather than having assessors rate a stimulus, the subjects' performance under stimuli of different qualities are being measured. Unfortunately, TPM results exhibit some undesirable properties because the do not always positively correlate with multimedia perception. Hearnshaw (1999) for example found that people achieved better results in remote teaching applications under poorer media quality because they increased - possibly subconsciously - their effort and paid more attention. Similarly, Reeves *et al.* (1993) showed that e.g. lower audio fidelity increased attention based TPMs. However, people were aware of the fidelity differences as reflected by their quality ratings. Thus, TPM do not provide a sounds basis for measuring QoE of passive content consumption, such as watching TV or listening to a song.

3.7.2 Physiological measures

Involuntary reactions can be obtained from users while they are being exposed to different media qualities to measure *user cost* or impact in terms of stress or arousal (Wilson & Sasse 2004). Examples for such measurements are galvanic skin response, heart rate, and blood volume pulse (BVP). Even though some people can learn to voluntarily influence these values (e.g. through feedback) data obtained from physiological measures are considered objective.

The tools to obtain these measures have become less obtrusive, and are being integrated into objects such as armbands and mice (Goulev 2004). A detailed automated interpretation of skin resistance, heart rate, and BVP requires the use of advanced statistical analysis and machine learning techniques (Mauss *et al.* 2004). However, the nature of the stimulus material (say, a live football match or an exciting thriller) can also have an effect on physiological measures; this can make it difficult to separate the effects of content from effects of video quality (Wilson 2006). Wilson & Sasse (2004) suggested that the use of several physiological responses in parallel (to e.g. eye tracking or micro-facial responses) may provide a way of disambiguating these effects. The data on changes in arousal can be enriched and interpreted differently with the knowledge of the location of visual attention.

Both physiological and task performance measures are not commonly used to measure perceived video quality or user experience but rather to measure the effect of video quality on human performance and user cost.

3.7.3 Eye-tracking

Current eye-tracking systems commonly use remote tracking technology, i.e. non-invasive techniques such that the tracking process does not hinder users. This is typically achieved by monitoring infrared light that is reflected from the surface of the eye. The reflection of this light differs as a function of the direction in which the eye is looking. Most retinal reflection systems are good at detecting the presence of an eye movement. The data can be used to compute the spatial and temporal distribution of gaze over an interface in terms of *saccades* (rapid movements to targets) and *fixations* (periods of relatively stable gaze during which the fixation target is perceived). Eye tracking is being widely used in usability research (Jacob & Karn 2004), (McCarthy 2003), but it also gives valuable insight into individuals' internal states. The measure *pupil dilation* provided by contemporary eye-tracking systems can, e.g., indicate *arousal* and *cognitive load* (Porter *et al.* 2003). The latency of saccades has also been identified as a measure of arousal (Roetting 2001). Eye tracking does not produce data that is totally objective since people have some control over their gaze patterns. It provides insight into what people are visually attending to and belongs to the category of observational methods. One drawback of eye-tracking in the domain of video quality asseassement is the fact that the human visual system is particularly sensitive to events that occur in the peripheral vision. Therefore correlating video quality ratings to the assessors focal area might yield inaccurate results as state changes in the periphery could be visually registered and affect video quality ratings.

3.8 Perceptual models

Barten developed the *square-root integral* (SQRI) to describe the effect of resolution on subjective picture quality. His extended model (Barten 1990) includes resolution, contrast, addressability, luminance, display size, viewing distance and noise. He showed that SQRI accurately models the subjective image quality results obtained by Westerink & Roufs, Jesty's PVDs (as the maximally achievable quality) and Mitsuhashi (1982). The latter is not available in the public domain.

A few models have been suggested to predict the perceived overall quality of audio-visual material models trying to quantify perceived quality, e.g. (Ghinea & Thomas 2001), (Hands 2004), (Winkler & Faller 2006), (Thang *et al.* 2007) and (Prangl *et al.* 2007).

Ghinea & Thomas suggested a model based on their definition of Quality of Perception (QoP). In their normative model QoP is a function (g) of two parameters $\mathrm{QoP}_{\mathrm{SAT}}$ and $\mathrm{QoP}_{\mathrm{IT}}$, which stand for the QoP of subjective satisfaction and of informational transfer. The latter denotes how well people pick up the informational content of the presented multimedia information. It represents a two-pronged approach with MOS and task performance as the building blocks for perceived quality:

$$QoP = g(QoP_{IT}, QoP_{SaT})$$

The performance of information transfer is measured by asking assessors questions about information that could have been obtained and was presented in one of the modalities: video, audio, and text in video clips. According to the authors QoP can be proportionally related QoS via their mapping formula, which distinguishes between three modalities: video *V*, audio *A*, and text *T*.

$$QoP_{IT} = \frac{V_{OK} + A_{OK} + T_{OK}}{V_{TOT} + A_{TOT} + T_{TOT}}$$

V_{OK} represents the number of correct answers pertaining to the visual information. In short, QoP is based on Mean Opinion Scores (MOS) and a task that measures information uptake by the user through three separate channels. Applying this measure in a study, Ghinea *et al.* found that a reduction of video quality from 25fps with a 24-bit colour depth to 5 fps with an 8-bit colour depth had barely any noticeable effect on either QoP_{IT} or QoP_{SAT} (Ghinea & Chen 2004) in an educational application setting.

Hands (2004) multimedia model considers both the influence of video quality and a multiplicative term of video and audio quality. According to his research multimedia quality (MQ) expressed in MOS depends on what is depicted i.e. the content and the shot types therein. For content with a shot type (head-and-shoulder, HS) akin to an MS both audio (AQ) and video quality (VQ) contribute equally:

$$MQ_{HS} = 0.17\ (AQ \cdot VQ) + 1.15.$$

For high-motion action scenes (HM) video quality becomes more important and was defined as:

$$MQ_{HM} = 0.25\ VQ + 0.15\ (AQ \cdot VQ) + 0.95.$$

Winkler & Faller (2006) suggest both a multiplicative and an additive model for audio-visual (MQ_{AV}) quality in MOS:

$$MQ_{AV} = 1.98 + 0.103\ AQ \cdot VQ_{.}$$

$$MQ_{AV} = -1.51 + 0.456\ AQ + 0.77\ VQ_{.}$$

Both Hands' and Winkler & Faller's model are based on subjectively measured MOS the ratings of which were obtained in subjective assessment situations that included anchoring. This limits the predictive capabilities of these models and especially its weightings to usage contexts in which the media qualities that users expect are within the quality range suggested by the anchoring.

3.8.1 Critique

The biggest problem of the given models is their lack of predictive capability. Video quality's poor correlation with people's viewing preferences has been shown in many studies presented in Chapter 2, e.g. (Pitts & Hurst 1989), (Lund 1993) and (Reeves & Nass 1998). Typical models for the VQ part in the perceptual models do not include content resolution, angular size of the depiction, aspect ratio or frame rate as variable input parameters. But all of these affect people's enjoyment of video content. The models are good at predicting what quality ratings assessors with sufficient anchoring and training will assign to a presentation but they are not indicative of users' preferences in terms viewing distances, aspect and viewing ratios both under lab conditions and in the field. Lund, for example, suggested that a *sense of reality* as induced by larger sizes might be more important to viewers than video quality. QoE models need to be able to predict these choices.

All current multimedia models are based on the assumption of additivity and/or multiplicativity. Higher quality in any dimension should result in higher overall quality. There are two possible ways in which users perceive inter-modal quality mismatches, i.e. what users experience when, for example, the text in the video is of better quality than the rest of the video. On the basis of *cognitive dissonance* theory (Feistinger 1957), one could argue that the video could be judged worse compared to a presentation without text because of the apparent mismatch between the two. Some studies on audio-visual interaction reported comparable effects when a quality reduction in the audio channel led to an increase in perceived video quality, e.g. (Beerends & de Caluwe 1999). Similarly, the perceived audio quality was judged of lower quality when it was presented with high quality video compared to the audio only condition (Storms & Zyda 2000).

The inclusion of task performance e.g. in Ghinea & Thomas' model poses other problems. As previously described participants compensate poor media quality by consciously or unconsciously increasing their efforts resulting in better performance but at the cost of higher stress and cognitive effort (Wilson 2006), (Reeves *et al.* 1993). In the context of Television consumption a number of authors (Chorianopoulos 2004) have pointed out that usability and performance measures do not adequately describe people's experience. Although informational transfer might be an important criterion for mediated information it is not clear if presentations that result in the highest informational transfer will equally achieve the best experience for the audience. Informational transfer is not understood well enough to be readily integrated into multimedia perception models. More is not necessarily better. In video conferencing blurred backgrounds are usually preferred by users although this conveys less information. Media production focuses on reducing the message and leaves out unnecessary information, which is not relevant to the story or the point that the creators want to get across. Still images or well-chosen key frames might be much better in conveying the information. The extreme would be a slide show of pictures. The medium of comics has shown that people's ability for closure from very few images is very good (McCloud 1994) and operationalizations of informational transfer such as recall might be better at very low frame rates.

3.9 Automated objective video quality measures

Automated objective quality measures aim at computing the visual quality according to a pre-defined metric. They do not require human assessors and should not be mistaken with objective measures obtained from human assessors such as TPM and physiological measure. Automated objective quality measures can be used to predict perceived image and video quality of individual clips. But advocates of objective measures agree that subjective assessments are more accurate and desirable for video quality assessment – unfortunately, with a higher price tag. The main reason for using automated objective measures is that obtaining subjective measure involves assessors, and data collection can require significant time and effort, which is impractical, e.g., for on-the-fly distribution of live content.

The most widely used image quality and distortion metrics are the purely mathematical *Peak Signal-to-Noise Ratio* (PSNR) and *Mean Squared Error* (MSE). More sophisticated approaches have tried to include models of the human visual system into their metrics (for example CVQE (Masry & Hemami 2004), DVQ (Watson *et al.* 2001)). In a comparative study conducted by the Video Quality Experts Group (VQEG) a wide range of metrics, which included PSNR, were tested. It was found that differences in their performances were not statistically significant (VQEG 2000). In a new approach Wang, Bovik, & Lu (Wang *et al.* 2002) are trying to capture structural distortion instead of errors with the *structural similarity index* (SSIM). However, in a recent update of the 2002 VQEG comparison study VQM (Video Quality Metric) by Pinson & Wolf (2004) was found to perform significantly better than the other objective quality metrics.

3.10 Discussion

The ITU MOS scale is very popular in audio and video quality assessment. However, there are number of problems with this categorical scale. Research has shown that the distances between the categories on the MOS scale are not conceptually equal in size (Aldridge *et al.* 1995), (Mullin *et al.* 2001). Similarly, research on photographic image quality showed that categories *excellent*, *very good*, *good, acceptable*, *unsatisfactory*, *poor* and *unusable* - very similar to the MOS scale labels were not equal in size (Corey *et al.* 1983) and therefore represented an ordinal scale. Critics of MOS dispute that the numerical labels alongside the categorical labels guarantee the interval property of the scale and question the validity of means computed from the numerical labels. Sporer pointed out that the repeatability of assessors' MOS scores from a five grade scale (with a 0.1 granularity) in audio assessment tests was quite poor and suggested that it might be useful to completely change the approach and use a binary threshold based approach (Sporer 1996).

According to Thurstone's model for scaling "Law of Categorical Judgement" – categorical values can be transformed into a linear scale (Torgerson 1958). If a repeated measures design is used in which many assessors provide categorical ratings of a certain condition and assessors rate multiple conditions a so-called class III, condition B model can be used. This transformation was for example done by Westerink & Roufs in their studies on subjective image quality of projected still images, which used the grades of the Dutch school system (from 0.1 to 10.0). However, most quality assessment studies employing MOS do not make use of Thurstone's transformation and simply average the numerical values.

Goodman & Nash (1982) tried to find out how well MOS scores of one site predicted those in a different country. From the results from seven different countries they concluded that MOS scores resulted in too large a variation and that further subjective tests would be required at each site. Furthermore, the ITU scale aspires to measure video or audio quality along a single axis but results indicate that neither of these two concepts are one-dimensional (Watson & Sasse 1998). In the domain of subjective assessment individual preferences within a population can result in e.g. binomial distributions. This means that aggregates as produced by e.g. MOS might not do any of the assessors' perceived quality justice as noted in (Sporer 1996). Despite all this criticism, MOS are widely used not least because of the widespread availability of mappings from objective measures to MOS, and the view in the quality assessment community that it is *'a standard scale*' that allows scores to be compared (Möller 2000).

3.11 Approach

As explained in section 2.6.8, video quality as operationalized e.g. by angular resolution does not provide good predictions of user preferences for experiencing video – e.g. in terms of their PVD. The objectively best video quality according to measurable fidelity and accuracy does not necessarily equate with what viewers like most. They usually prefer more vibrant colours that are not necessarily accurate or natural (de Ridder *et al.* 1995). Pictures with the highest amount of detail i.e. sharpness are not perceived as having the highest quality. Beyond a certain point, people find an increase in sharpness disagreeable and it results in a lowered perceived quality (Frieser & Biedermann 1963). Common to all these finding is that people's preferred visual experience does not equate to the highest visual quality. HDTV was designed to provide better immersion an experiential factor different from video quality. The original contraption to measure HDTV immersion was very sophisticated and subsequent research has not replicated these setups but have often chosen to operationalize QoE through video quality.

In my research I made use of a range of methods to assess the QoE of a service. The participants' instructions stated that the goal was to find out which presentations they found acceptable for a specific service. I used the method of acceptability because together with the framing it did not preclude any experiential dimensions that people might deem necessary from their ratings. It included contextual factors - the participants watched realistic content they were interested in on actual mobile devices at typical viewing distances. The results provided an easy interpretation in terms of QoE for a population of users. Binary measures are easy to understand and they reduce the cognitive effort in the decision that participants have to make. This in turn increases the reliability and lowers the variability of the ratings that are typically high in categorical measures. For most experiments the ratings were not called out but the participants provided their ratings by means of a visual interface (*cf.* Figure 19, page 99). It allowed for continuous intra-stimulus ratings in video clips of realistic length. This approach should yield ecologically valid results and be open to long term effects if existent. The analysis of binary measures is straightforward and does not involve further mapping as required for the correct use of categorical measures. Finally, the low overhead makes them easy to use in non-laboratory situations as in study 3 on the train.

This quantitative data gathering was combined with qualitative feedback obtained through debrief interviews in which the participants commented on their ratings and provided reasons as to what prompted them to assign their ratings. The feedback obtained allowed for a more in-depth understanding and validated that ratings were not solely based on video quality but on other factors such as, for example, the size of the video. The combined approach therefore integrated the assessment of the visual experience constructively into the human-centred design approach as the ratings were obtained in the context of the overall experience provided by multimedia content presented under realistic conditions rather than a focussing on video quality alone.

Chapter 4

Many a small thing has been made large by the right kind of advertising.
- Mark Twain

Mobile multimedia consumption

Mobile devices are becoming an increasingly popular means to consume and interact with multimedia content. I carried out my research at a time at which many digital technologies based on different standards had started to enable different consumption paradigms. Delivery of the content ranged from traditional broadcasting, over multi- and unicast to media charging and physical media. This chapter presents the background on mobile TV research. It reviews the extant research findings on the context of mobile multimedia consumption including the location and social context, the motivation of use and the interaction with the content. It includes the different interaction paradigms and technological solutions that enable this activity. It reviews the key dimensions for viewing on mobile devices along with the trade-offs and adaptations that have been pursued to enhance the user experience.

4.1 Introduction

The media landscape is in a state of flux. Digital TV is becoming a reality and analogue TV will be switched off in many countries in the not too distant future. Large high-definition resolution screens are becoming more popular for a home theatre experience. Historically, portable TV sets were the only way to watch TV anywhere but for a mobile device they were large and consumed a lot of energy. Portable television sets have, over time, decreased greatly in size, but until now, TVs did not become a mobile gadget that many people carry around with them. Portable entertainment hardware has evolved, from portable radios and walkman to mobile gaming consoles, portable music players that boast more storage than the average digital music archive of consumers and mobile phones that have almost usurped personal digital assistants (PDAs). Most of them are now capable of rendering video content and can provide TV-like experiences. They vary in size, energy consumption and follow different content delivery and consumption paradigms. Content can be either played back from storage (e.g. DVDs) or device memory (e.g. iPod), delivered on demand by mobile operators, received live by broadcasters or downloaded through the Internet, streamed from computers relaying broadcast and stored content (e.g. Slingbox) or set top box solutions (e.g. AppleTV). This shift encompasses not only the delivery but also the production and augmentation of content. The barriers to entry for content production are dropping meaning that non-professional can produce or edit content with increasingly less effort, and traditional content producers are therefore not the sole sources for appealing content.

4.2 Brief history

Since the wide availability of TV receivers after the Second World War, the cinema industry has time and time again augmented the experience in the theatres from mute black and white pictures by introducing sound, colour, widescreen aspect ratios, high fidelity and surround sound to stay ahead of the experience offered in the home through the TV. In the 1980s, Seiko introduced a TV wristwatch that was capable of displaying standard TV channels on an LCD wrist watch. A growing number of people used LCD or digital watches, and it was possible to display rasterised images on an LCD display. It had a monochrome screen with 10 shades of grey in a bluish tint with a small screen diagonal of 2.8cm and 31920 pixels (210x152). However, the watch was not a success. One of the biggest problems was high energy consumption - the watch wearer had to separately carry the 2AA batteries, which was part of a box that housed the TV receiver and connected to the watch through a cable – a restrictive and finicky setup. This setup limited the wearer to a few hours of viewing time. The screen was monochrome and had low contrast with no backlight. Watching TV, while wearing the watch, resulted in an un-natural wrist posture. Last but not least, the TV wristwatch was expensive. Twenty years after the wrist watch TV a number of technologies have emerged that allow for the consumption of audio-visual content without the historical restrictions. People now carry inexpensive mobile phones, music players and game consoles with built-in screens. This allows the display of moving images. The distribution of content is possible in more energy efficient ways, too.

Both broadcast and unicast Mobile TV services on mobile phones are available in a number of countries now. A few studies have trialled mobile TV broadcast services, e.g. the BT movio (Lloyd *et al.* 2006) and the Oxford study (Mason 2006) in the UK, the MiFriends study in Germany, Nokia in Spain (Nokia 2006) and the VTT study in Finland (Södergård 2003).

4.3 Mobile multimedia experience - context

The fact that video recorders (VCRs), DVDs and personal video recorders (PVRs) have not forced cinemas out of business might have different reasons. The increase in the working population's free time surely helped, but it is more plausible that the experience of watching a movie in the cinema is different enough from watching at home. This is not only due to the higher audio-visual quality and immersion, but also the context of the location, social setting and protocol and personal dedication, which make for a different experience than watching the same movie possibly adapted to the TV screen ratio, in the living room under inferior lighting conditions with a variety of sources for interruptions present. TV represents a very different medium from cinema and it is quite likely that mobile TV, if successful, will be medium distinctly different from mobile cousin of TV, but possibly something rather different – a personal technology, yet interactive and fostering communication. The following subsections explore this potential by reviewing previous research on important contextual factors that shape mobile multimedia consumption: where does it take place, who else is present, why do people use it and how?

4.3.1 Location

A number of contextual factors that are associated with the space in which a technology is used might affect people's experience of a mobile multimedia service. Factor such as ambient noise, lighting might make it harder to enjoy multimedia if it is possible in terms of coverage, reception and power constraints. Mäki *et al.* (2005) identified the home, work and transit as the three top places for mobile TV. Between 30% to 50% of the participants used mobile TV at home in different mobile TV field trials (Mason 2006), (Lloyd *et al.* 2006), (Nokia 2006). In Södergård's (2003) study people used PDAs and tablet PCs in the home context suggesting that screen size does not affect this choice much even when larger screens with higher resolution are available. Chipchase *et al.* (2007) observed that mobile TV constitutes rather a personal TV which is therefore used in the home, too. Mobile multimedia services that are neglecting indoor coverage might fall short in terms of how people want to make use of mobile services. Research has shown that people feel restricted by a lack of screens in the home (Seager *et al.* 2007). They want to use multiple entertainment, communication and information services in parallel and possibly in different parts of the home. In order to achieve this people have started to appropriate laptops and other mobile devices to satisfy their needs. Mobile devices can act as additional screens, remote controls and display additional, interactive information (Cesar *et al.* 2008).

4.3.2 Social context

The social organisation of households plays an important role in how technology gets used. In their ethnographic study O'Brien *et al.* (1999) found that the concentration of functionality (TV, Internet *etc.*) in the living room through set top boxes (STB) did not allow for a natural distribution of activities across different people and spaces in the home. Hughes *et al.* (1998) pointed out the relationship between technology use and ownership of space. People make a claim on the space in which they are using entertainment systems. The use of television or stereo systems indicated a control of space that might not be intended in every case but might be a by-product of convenience and the limitation of the existing technologies. Koskinen & Repo (2006) found that people moderated their use of mobile TV devices in shared spaces to save face. They used it either politely to avoid disturbing others and adjusted this further in case of disapproval or aggressively used the devices to draw attention.

Mobile devices especially when equipped with ear- or headphones do not impose the ownership of space as described by Hughes *et al.* In that configuration these devices make for a *quiet technology* such as text messaging (Grinter & Eldridge 2001) which does neither create or require sound that might disturb others in a shared space. They provide a more granular control of space. This could be a driver especially for younger users that do not have a large say in the use of shared space in the home. Mobile devices allow for a straddle of media use and shared space as well as a choice whether or not to conform to social norms while watching.

4.3.3 Motivation of use

Peoples' watching of standard TV is driven by ritualistic (Taylor & Harper 2002) and instrumental motives (Rubin 1981) as in '*electronic wallpaper*' (Gauntlett & Hill 1999), mood management (Zillman 1988), escapism, information, entertainment, social grease, social activity, and social learning (Lee and Lee 1995). For many of these drivers watching TV constitutes a group activity. Whereas the drivers behind standard TV consumption are fairly well understood, comparable knowledge in mobile TV is lacking. A number of studies focused on people's usage of mobile devices for the consumption of video - see (Harper *et al.* 2008) for an overview. Södergård (2003) and Repo *et al.* (2004) reported both individual and collective viewing on mobile devices. O'Hara *et al.* (2007) and Chipchase *et al* (2007) provided studies on how and why people consume video material on mobile devices. O'Hara argued that even though consuming video on mobile devices is a privatizing technology, it might facilitate togetherness in the home as people can watch "*their own content while being in proximity to family*".

Harper *et al.* (2008) pointed to the active and social component of TV watching on mobile phones and used the term *watching to show* as the maxim to describe this salient property.

4.3.4 Interaction with content

In many industrialised countries people watch TV for an average of three hours a day. Södergård (2003) found that spurts of watching time on mobile devices were quite short in comparison to regular TV - between 2 and 5 minutes, and that news was the most demanded content class by all user groups. Yanqing *et al.* (2007) argued the time required to set up the device in a given context was favouring use in macro- rather than micro-breaks of a few minutes.

Although usually described as a passive, lean back activity people exercise a fair amount of choice while watching TV. Organization of multimedia content into channels that can be easily switched between is the predominant interaction paradigm in broadcast services, but it is not common on media chargers and playback devices. Taylor *et al.* (2002) argued that channel surfing is inherently associated with the act of watching TV. The methods to select a program used in traditional TV viewing depend on the time of day. But the method used generally escalates – if nothing of interest is found – to strategies that require more effort on behalf of the user. The order of strategies is:

(1) channel surfing,
(2) wait or search for a TV program announcement,
(3) knowledge of weekly schedules or upcoming programmes,
(4) paper-based or onscreen guide.

Channel switching in on demand and broadcast services constitutes a challenge in terms of user satisfaction. There exists no research on users' wait time tolerance for leisure multimedia consumption on mobile device or in mobile contexts but the current channel switching delays of around three seconds and more in many services, was regarded as the biggest usability problem in mobile TV services at the World Handset Forum 2006 in San Diego (Weiss 2006). Miller established the rule of 2 seconds for human-computer interactions as an upper limit. If commands issued by the users do not result in some feedback of the system within 2 seconds the interactions lose their conversational nature and the ideal 0.5 seconds results in the highest "conversational" flow between human and computer. After longer waits, i.e. 10 seconds users get "distracted" and might move on to another task (Miller 1968). Long waiting times after a requested channel switch will result in lower user satisfaction. Since users are accustomed to almost instantaneous switches on standard TV, the delay should be as short as possible. First results for digital TV indicated that 0.43 seconds might be the limit beyond which users will be increasingly dissatisfied (Ahmed *et al.* 2006).

In digital TV, the switching delays depend to a large part on the video codec, e.g. in MPEG encoded content on the occurrence of so-called key frames. Fewer key frames in a video broadcast result in smaller amounts of bandwidth required to transmit the content but the receiver has to wait for the arrival of the next key frame in order to be able to display a newly selected channel. Service providers could exploit the fact that the human visual system is inert. An average recovery time of 780msec between scene changes was acceptable to even the most critical observers, when visual detail was reduced to fraction of the regular stream (Seyler & Budrikis 1964). Further research would be needed to see if this period applies equally to channel switching on mobile devices, and how codecs could make use of this period.

Design tricks can be employed to reduce the perceived wait time, e.g. - displaying the logo of the upcoming channel, or playing pre-stored material advertising other shows. Long wait times for downloading or on-demand streaming content should be accompanied with progress bars to help users assess the remaining time (Serco 2006). A number of approaches have been devised to reduce the start-up times of Video on Demand (VoD) services, which try to improve bandwidth utilization while achieving short start-up delays. In *batching* the server waits for a number of requests for the same clip and then starts multicasting them (Dan *et al.* 1994). In Buchinger & Hlavacs' (2008) *low start approach* people receive video on demand content instantaneously in but at a lower quality at the start, which increases once the content can be received through multi-cast.

4.4 Provisioning of mobile TV

Many different solutions exist for people to enjoy audio-visual material on mobile devices. The group of *play-back* devices comprises portable DVD players and portable play stations that require physical media from which the content can be read. *Media chargers* store content on their local memory and typically obtain it through the Internet or home entertainment components such as set-top-boxes. *On demand* (unicast) services deliver the content to a mobile device through a cellular network to individual customers. Services that *broadcast* content to all switched on devices make up the last category. Playback devices and media chargers do not deliver the content over the wireless spectrum but since they enable mobile multimedia consumption I have included them. A number of broadcast and unicast standards have evolved in the field of mobile TV, see Kumar (2007) for an overview. The most salient properties of these four mobile TV paradigms are listed in Table 7. *Geographic, legal* and s*peed restrictions* refer to the fact that wireless services may be limited due to reception; that legal restrictions might curb consumption outside jurisdictions; and that reception imposes limits on the speed at which the receiver can move. Since reception might vary in terms of strength and quality streamed and broadcast services might not be able to *guarantee continued presentations* of content and the presentation might stall at times. The audio-visual *quality* of the content is higher when the user receives the content through non-wireless distribution mechanisms. Live or near-live content such as football games or newscasts is typically inaccessible through playback media or is accessible only with large delays on media charger solutions – the *recency* of such content is limited. The *programme range* denotes the amount of content that can be made accessible. The *interaction paradigm* describes how people can choose from an array of programs. The *delivery* method refers to how the content is distributed either to people individually or via a broadcast mechanism. The decision maker of which content can be consumed at a given time is denoted by *decision of content. Scalability* refers to whether the consumption paradigm can easily satisfy a growing user population.

The wireless domain is one of limited bandwidth resources, and service providers have to decide on broadcasting more content at lower quality or vice versa in search of optimal configurations for people's QoE that are financially viable. In analogue television the trade-offs were different. The available frequency bandwidth had to fit the spatial, temporal, colour and luminosity information. Once the division between these parameters was standardized any innovation had to ensure backwards compatibility as was the case for the introduction of e.g. colour and tele-text information. The parametrical space of digital mobile services is more complicated. This is exacerbated for non-broadcast services in which service providers have an incentive to further optimize the individual delivery to devices. When adapting to mobile devices, service providers face a range of target display configurations in terms of aspect ratios, sizes and supported codecs. There are three main ways to address the problem of addressing target devices of different resolution:

1) broadcasting at the highest resolution and resizing on the receiver side,
2) sending multiple resolutions, which requires more bandwidth if broadcast or
3) employing layered coding schemes that broadcast a number of resolution layers from which every receiver can assemble the parts it can display. In the H.264 video coding standard this technique is referred to as scalable video coding (SVC) (Richardson 2003).

Table 7: Mobile multimedia consumption paradigms and limitations

	Play-back	Media charger	On demand (unicast)	Broadcast
Example	mobile DVD player	iPod	3G TV	DVB-H, T-DMB, mediaflo
Geogr., legal & speed restrictions	none	none	within network cell at designed speeds	in reception area, at designed speeds, better outdoors*
Guaranteed continuity	high	high	medium	low
Quality	highest	high	low to medium depending on cell	medium
Currency	low	medium - since last contact with base or delivery network	high	high
Programme range	large but not any time	large	large	small
Interaction paradigm	select from media	select from storage	select from list or virtual channel hopping	channel hopping or EPG
Delivery	individual	individual	individual/group	everybody the same
Decision of content	long term decision, by user	a priori, by user	on demand from range, by mobile operator	live, by broadcaster
Scalability	high	high (e.g. p2p)	low	high

* Currently, most broadcast solutions are aimed at outdoor coverage (Yoshida 2006).

4.5 Viewing experience

This section extends the dimensions introduced in Chapter 2 with research based on mobile devices.

4.5.1 Screens

During the 1990s flat screen computer displays started replacing CRT displays but mobile devices such as watches and household appliances had been using liquid crystal displays for a long time because of size and energy constraints. This has made LCD displays an obvious choice for battery driven mobile devices such as laptops, mobile DVD players, PDAs, mobile phones and other appliances. Lately Organic Light-Emitting Diode (OLED) displays have been making advances and allow for brighter, more energy efficient and wider viewing angle displays. Due to current size and luminance constraints projectors have not seen much application in mobile entertainment but there are already existing projectors – albeit with relatively low luminosity - that could be built into mobile devices. In the beginning mobile devices were limited in terms of rendering video content in terns of resolution and colour depth limitations due to computational ability to decode and the ability of LCD screens with their restrictions in refresh rates, resolution, contrast and luminosity. Nowadays mobile devices with 200ppi and VGA resolution an higher exist that exceed the targeted broadcast resolution of mobile TV in DVB-H (QVGA) and that can render content at SDTV resolution. However, the mobile device landscape's diversity makes delivery of multimedia content a challenge for services based on media chargers and on-demand (unicast) systems. According to a study by Serco (2006) landscape oriented use of the display might be preferred over the typical portrait mode that mobile phones are used in. Recently a number of devices have introduced swivelling screen that can be used in both landscape and portrait mode.

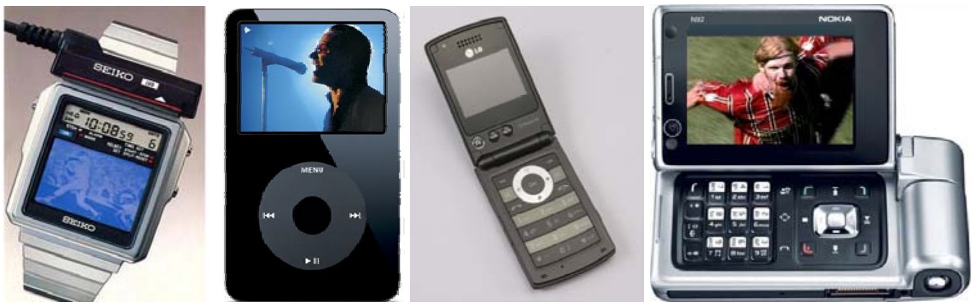


Figure 14: Mobile TVs (l.t.r.) SDTV watch, digital video player, DVB-T and DVB-H phone

As long as screens cannot be folded away or rolled out, the size of the device as a portable medium remains a concern (*cf.* Sec. 5.2.4). Projectors, head mounted displays and auxiliary screens may be possible solutions to watch content but are not adopted for this purpose yet. Recently, airlines have announced to provide connections for iPods to use the backseat displays. Projectors will become smaller, more energy-efficient and more powerful, and able to project in regular daylight. Head-mounted displays will become smaller and lighter, and integrated into glasses. Since mobile devices are not always exclusively used for content consumption the video might have to share the available screen estate with other information, e.g. in services like co-viewing, that allow for chatting while watching TV.

4.5.2 Viewing distance

Paper, keyboard and display objects are typically operated at distances ranging from 30cm to 70cm. Personal displays fall into the same category. Lund predicted a PVD of 53cm for very small picture sizes. Fatigue and stress are valid concerns for continuous watching at a distance closer than the resting point of vergence - approx. 89cm, with a 30º downward gaze. This is a posture often seen in mobile TV consumption (*cf.* Figure 6 and Figure 15) because people use their legs or bags on their laps as support (Yanqing *et al.* 2007). Kato *et al.* (2005) obtained typical viewing distances of approx. 35cm from both standing and sitting people using a 166ppi mobile device at 11*H*. Although the study did not address possible effects of resolution the results from Kato *et al.* confirm that mobile TV will be consumed at around arms` length but the PVD has not been researched in relation to size and resolution.

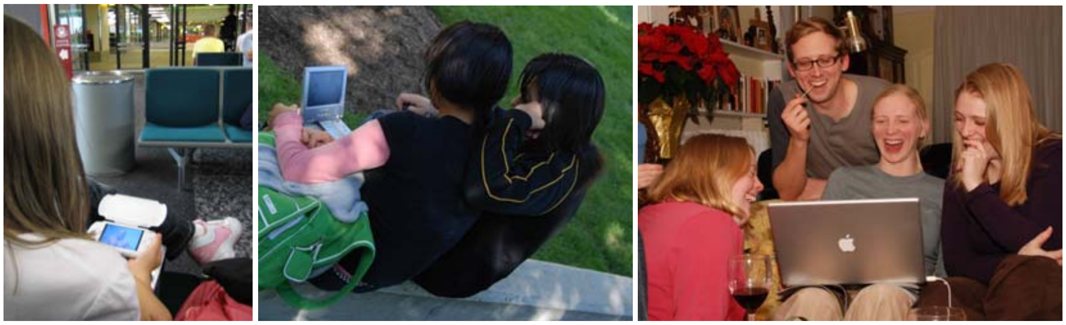


Figure 15: Mobile viewing ratios varying (from left to right) from around 8*H* to 1.5*H*

Viewing distance has an effect on the possible social uses of the screen. In general, short viewing distances cannot accommodate as many viewers to share a screen in comparison to longer viewing distances even at a constant VR. This is further exacerbated by the fact that the angle from which a screen can be viewed is limited and portable devices tend to have smaller *viewing angles* than the more powerful TV sets. On the positive side, smaller viewing distances allow for more private content consumption. Head mounted displays (HMD) and retinal projectors afford very little opportunity for peeking and can therefore guarantee exclusive viewing. For these displays the viewing distance is not the absolute distance between the visual centre of the eye and the device but the distance at which the eyes have to converge and accommodate in order to obtain a sharp image.

4.5.3 Content

Professional content is produced with a specific primary target medium i.e. cinema, TV or mobile in mind. This choice influences the selection of shot types, length and type of the programme. In comparison to cinema content, which takes a long time to produce, TV content production is fast. TV programmes are typically shorter in comparison, lasting less than 60 minutes and geared to be presented daily or weekly. A notable exception are live broadcasts, e.g. of sports or other events of common interest which have longer durations. News is produced mainly for TV consumption. Producing bespoke mobile content is expensive, and most customers are reluctant to pay a premium for mobile content (KPMG 2006). Thus, service providers look for automated, low-cost solutions to repurpose existing material and maximize the user experience on mobile devices – ideally simply recoding existing TV or cinema content.

4.5.4 Content adaptation to mobile devices

Some content producers tailor make content with respect to low resolutions and short viewing time, e.g. abridged versions of the popular TV series 24. Germany's public broadcaster ARD started producing news programmes for mobile consumption that are 100 seconds long and contain large text. MTV is re-subtitling their content for mobile consumption with larger, shorter and sharper fonts made for mobile (Kelly 2006). Text was a large factor in perceived video quality in Jumisko-Pyykkö (2007) study on mobile TV consumption.

In Asia, content creators produced soap operas for mobile devices, which are short and rely heavily on close-up shots. Most emotions have to be conveyed by means of facial expressions and *"there is very little dialogue and a lot of close-ups of characters striking exaggerated poses"* (Guardian 2005). In sports coverage for mobile devices ESPN minimises the use of *long shots* in their coverage (Gwinn & Hughlett 2005), and uses more highlights with close-up shots instead. Unfortunately, due to inconsistencies in shot type naming it was not clear whether the article referred to *long shots* as described in Sec.2.3.9 or shots with even less detail (VLS and XLS). Content-based pre-encoding techniques can improve on the visual information and detail by:

(1) Cropping off the surrounding area of the footage that is outside the final safe area for action and titles and does not include essential information. The broadcast material includes this to compensate for maladjustment of TV receivers (Thompson 1998).

(2) Zooming in on the area that displays the most important aspects (Dal Lago 2006), (Holmstrom 2003). This means that a given shot type is being transformed into a more detailed one (*cf.* Sec. 2.3.9). The resolution of the original content imposes limits on this approach.

(3) Highlighting and increasing the size of the object that yields the highest amount of interest, e.g. the ball in football or the puck in ice-hockey. While increasing the contrast of the ball had a positive effect on the user experience, increasing the size of the ball to enhance recognition made the presentation worse unless the video was encoded at very low bitrates, e.g. 28, 32kbps (Nemethova *et al.* 2004).

However, the gain of these changes is not fully researched or understood. No published reports were available on the influence of low resolutions on the different shot types used in television content and how these would come across on mobile devices. Research is required to rule out possible negative side-effects caused by these automated approaches.

4.5.5 Zooming

The most obvious solution to increasing the amount of detail is to zoom in on part of the material and crop off the remainder. Zooming can be done in with a static and a moving window approach.

Static window: TV content is produced so that misalignments of the receiving analogue TV sets do not impair the viewing experience. The content contains a so-called safe area outside of which no important information should be presented (Thompson 1998). Static cropping could therefore zoom in on the safe area without omitting important information.

Moving window: Zooming in on the area displaying the most important aspects (Dal Lago 2006) (Holmstrom 2003). This is similar to the pan-and-scan approach when presenting wide-screen cinema footage on 4:3 TV screens without black 'letterboxing' bars. There are number of steps that need to be taken to apply zooms. Identification of shot boundaries, classification of the shot types, detection of the ball, identifying the region of interest (ROI) and avoiding oscillations of it in adjacent frames to avoid disturbing jitter effects as pointed out in (Kopf *et al.* 2006), computing the acceleration, deceleration of the zoom and the panning speed of the moving zoom window. In (Agarwal *et al.* 2003) this was done algorithmically by calculating the ROI based on the human visual system (HVS). Another solution entails employing a human observer to make the decision of zooming on-the-fly by means of eye-tracking technology (Agro 2005). Moving windows introduce additional *secondary motion* on top of the pans and zooms of the original footage. This additional panning needs to be controlled because viewers object to both sudden jumps as well as excessive panning in the footage. The latter has been likened to drunken camera operators (Holmstrom 2003). On top of this the amount of zoom and the area that is cropped off can be dynamically adjusted. In (Agarwal *et al.* 2003) this dynamic zooming approach was compared to an approach with a fixed zoom factor. Dynamic zooming was identified as superior to the fixed zoom but the paper failed to report how the tests were conducted and how many participants were involved.

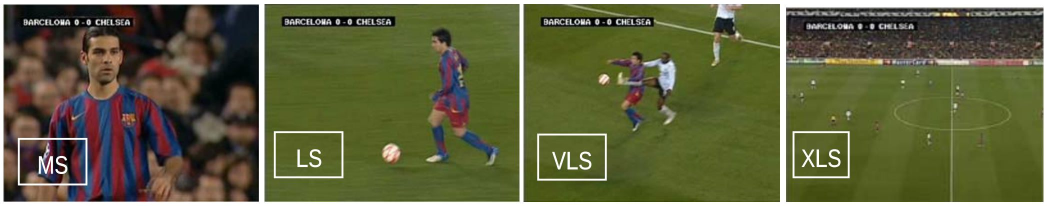


Figure 16: Shot types used in sports coverage from left to right: Medium shot (MS), long shot (LS), very long shot (VLS) and extreme long shot (XLS)

The multimedia research community has embraced the idea of zooming into pictures to improve the viewer's experience on small screens e.g. (Kopf *et al.* 2006), (Sinha & Agarwal 2005), (Seo *et al.* 2007). But a number of concerns remain about the range of zooms that can and should be used. To date, none of the zooming techniques have been evaluated on mobile devices. Most importantly, it is unknown how much zoom is advisable for which source resolution, target size, target resolution and shot types. Whereas the benefits at first glance may seem self-evident, it is not clear that the perceived gain in visual detail will outweigh the contextual information lost due to cropping, or whether that information is necessary to understand the context of a scene. This may be especially true for field sports, such as football: the extreme long shots that make up the majority of this content cover a large amount of the pitch, and the audience can benefit from seeing potential pass receivers or other strategic information.

Another problem with passive viewing scenarios is that people do not necessarily want to interact with the content in order to actively initiate or control a zoom. However, broadcast content needs to be adapted to the different resolution and screen sizes of mobile devices in use. Default values for zooming for the different resolution could help here. The benefits of zooming into content using extreme long shots will be investigated in Chapter 10.

4.6 Multimedia trade-offs on mobile devices

Content adaptation might be necessary when content needs to be presented in a medium different from the one for which it was produced. An overall trade-off between quantity and quality dominates the delivery of content for service providers. The lower the quality of content the more content can be delivered making a service more attractive to more customers. If the quality is too low people will not pay for and use the service. To maximise their profits service providers need to find the optimal combination of volume and audio-visual quality of the content. The latter depends primarily on how much information (in bits) is used to encode and represent the media. This overall budget can be allocated differently to the audio and video components. For both media there are yet more decisions to make. For example, how much information should be used to encode the temporal and the spatial information in terms of frame rate on the one hand and amount of pixels, quantization and colour depth on the other hand. To some degree angular size and resolution can be traded-off by the viewer by adjusting their viewing distance or on some devices by increasing the image size.

4.6.1 Audio-visual interaction

Based on Kaasinen's (2005) application of TAM to discretionary use for mobile services, Jumisko-Pyykkö & Väänänen-Vainio-Mattila (2006) claimed that audio-visual quality affected users' trust towards a system, and finally the willingness to use mobile services. A number of studies have found that the combined quality of audio-visual displays is not simply based on the sum of its parts e.g. (Hands 2004) and (Jumisko-Pyykkö & Häkkinen 2006). In a study on audio-visual interactions, Winkler & Faller (2005) found that selecting mono audio for a given video encoding bitrate gives better quality ratings than stereo and that more encoding bitrate should be allocated to the audio for more complex scenes.

4.6.2 Spatio-temporal trade-off

Spatial and temporal resolution are key factors for the perceived quality of video content. Whereas temporal resolution below 30 fps results in successively jerkier motion of moving objects, lowering the number of pixels to encode the picture reduces the amount of visible detail. The higher the resolution along both of these dimensions, the more bandwidth is required for transmission. Bitrate constrained encoders typically use a mixture of temporal and spatial resolution reduction when trying to meet the requested bitrate constraint. Excessive delays and loss of content during transmission may affect both the spatial and temporal resolution resulting in visible artefacts like blocks and/or skipping of frames causing the picture to freeze. Although it is very useful to be aware of the individual boundaries explained in Chapter 2, it is even more important to select good combinations of frame rate and resolution.

Kies *et al.* (1996) conducted two studies to determine the effect of reduced video image quality on task performance and subjective preference in a distance learning application. The first was a controlled laboratory experiment performed at different frame rates (1, 6, 30 fps) and resolutions (320x240, 160x120) targeting the subjects' performance in a quiz. The second was a field study set to better understand the effects of reduced video image quality in a natural, realistic setting with motivations, concerns, and factors associated with a real class. The results from the performance-based experiment did not show significant effects for a decreased frame rate and/or resolution. The evaluation of questionnaires in the second experiment recommended avoiding frame rates less than 6 fps and resolution settings less than 320x240. However, it the publication does not specify how the video was presented. Pappas & Hinds (1995) examined the trade-off between frame rate and level of greyscale for viewing a pre-recorded video segment of a brief technical lecture. They found that when frame rate dropped below 5 fps, subjects were willing to make tremendous compromises to avoid these low frame rates, e.g. choosing binary grey level over full grey scale.

Assfalg *et al.* (2003) and IBM (2002) recommended prioritizing frame rates for fast moving content such as football, over spatial resolution but this was not based on empirical research. A number of studies showed that this content was not very sensitive to frame rate changes (Apteker *et al.* 1994) and (Ghinea & Thomas 1998). Wang *et al.* (2003) reported on a study in which they manipulated both frame rate and quantization with an American football clip. They concluded that *"quantization distortion is generally more objectionable than motion judder"* and that large quantization parameters should be avoided whenever possible. Hauske *et al.* (2003) used resolution typical of mobile devices but presented them on a monitor at 6*H* (27cm viewing distance) assessors rated low frame rate (between 15 and 5 fps) and low bitrate (30-50 kbit/s) QCIF resolution video clips using absolute category ratings (ACR). Participants were anchored with a clip at 128 kbit/s with 15 fps. It turned out that video quality was more based on blocking effects and information value than on the smoothness of movement. Similarly, people preferred higher spatial resolution over higher frame rates in order to be able to identify objects and actors in football content on mobile devices and computer screens (McCarthy *et al.* 2004a). The only study in support of prioritising frame rates over resolution was Song *et al.* (2004). They had addressed the display of content in mobile environments comparing content encoded at two bitrates (348kbit/s and 1.5Mbit/s), two resolutions (174x144 and QVGA 320x240) and three frame rates (5, 15 and 25fps) under no packet loss. The participants had to rate the QoS of the video quality under different frame rate, frame size and playback bandwidth. The authors identified frame rate to be the most influential factor on ratings. The study failed to explain what it means "*to rate the QoS*" of the video quality and how the clips were displayed, which makes the interpretation of the results difficult. But the participants might have simply equated low frame rates with jerky video playback, which is common in low QoS conditions and would resolve the contradiction to the results of McCarthy *et al.*, Hauske *et al.* and Wang *et al.*

Although individual parameter effects are important they have to be considered together in order to achieve the best possible QoE. The trade-off between resolution and frame rate is not straightforward as the above studies might suggest. Encoders are typically bound by a target bitrate and will aim to achieve a given target frame rate for a nominal spatial resolution. This may entail a reduction in spatial definition.

4.6.3 Encoding bitrates and resolution

HDTV required five times the bandwidth of SDTV to achieve a larger picture with more detail at the same angular resolution. Digitally encoded content requires a fraction of the bandwidth of its analogue counterpart in distribution. However one of the fundamental questions for video encoding is the amount of encoding bitrate required to achieve a certain level of quality for a given target resolution when other factors such as frame rate are kept constant. Just because the target resolution is, e.g., VGA it does not mean that the depicted content includes spatial information that requires it. The same is true in the temporal domain. In the spatial domain high frequency spatial information such as edges e.g. grating patterns, text and small objects require high resolution and depending on their suitability to be compressed end up requiring more information to represent them. Figure 17 illustrates this trade-off. If fewer pixels require encoding the encoding bitrate per pixel is higher than for more pixels and should result in a better visual quality. But what is the optimum resolution for a given bandwidth and target display size? In Westerink & Rouf's study both size and resolution independently affected the resulting visual quality. For broadcast channels a typical service configuration mentioned in (Mason 2006) and (Lloyd *et al.* 2006) is QVGA resolution at 25fps

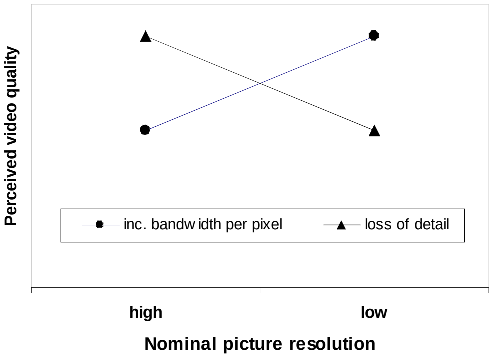


Figure 17: Trading off nominal resolution for higher encoding bitrate per pixel

encoded at 250kbit/s but display size is not considered as a factor.

4.7 Summary & conclusions

The following are the core findings that stem from studies that have been obtained on mobile devices directly or are based on studies aiming at simulating multimedia content consumption on mobile devices

1. The lowest temporal resolution to be acceptable for people to follow video is 5 fps but on mobile devices already depictions lower than 12.5 fps resulted in degraded experiences. At the same time resolution seemed to be more important for the acceptability of video quality. I therefore considered 12.5 fps as a lower bound and used it as a nominal value in the generation of video clips in all experiments described in this thesis in the following chapters.
2. Viewing ratios can range from 20H to 1.5H depending on the mobile device's screen size. This might result in problems for text legibility, shot types and required resolution of the content.
3. Many previous studies did not include extreme long shots (XLS) but relied mostly on shots in the range between medium close-ups and long shots, which show much more detail - shot types posed a problem in HDTV research and current content production.
4. A single study on viewing distances on mobile device suggests that it might be considered fixed.
5. Usage both in private and in shared settings will be common – especially indoors.
6. Current multimedia models do not provide predictions of user preferences for mobile consumption of moving images or audio-visual material. The model of Barten that includes both size and resolution concerns was based on Westerink & Rouf's results on still pictures, which were rated based on visual quality not on preference. Their optimal point of 32 ppd can be helpful as a starting point for further research but might neither provide us with settings preferred by the user nor result in the highest QoE.
7. Although early studies by e.g. Jesty suggested that viewing distance could be used as an indicator for visual quality more research in a larger parametrical subspace by Lund showed that visual quality was not a reliable predictor for preferred viewing ratios for TV consumption.

In the following chapters I will describe the studies that were designed based on the current knowledge presented in this and Chapter 2. Taken together with the results it can form part of an empirically grounded understanding of QoE in mobile multimedia services that can further multimedia perception models and guide practitioners in the design of financially viable services.

Chapter 5

Preparatory studies

In parallel to the literature review presented in Chapter 2, 3 and 4, I was involved in carrying out an exploratory study to identify factors determining QoE in mobile multimedia applications, specifically a mobile TV service. In 2004, interview partners who had experienced a mobile TV service first hand were difficult to obtain, since the only operational services were in Korea or Japan. I conducted a series of focus groups in the UK and supervised many that were facilitated in France by a native speaker. The goal was to find out more about people's current experiences with mobile phones and services, as well as their future needs and expectations of mobile TV services hosted on mobile phones. This research was carried out within the requirements phase of the MAESTRO project (Selier & Chuberre 2005).

5.1 Interviews

Interviews with two people from South Korea, who had previously experienced mobile TV services first hand, were used in addition to the focus groups. The interviews provided a first glimpse of a more grounded understanding of views and problems with mobile consumption of multimedia content. The interviews were conducted in a semi-structured way, which allowed respondents to elaborate on any points they felt important (Breakwell 2000). The interviews suggested that one of the biggest problems of the existing mobile television services was cost and lack of transparent billing. The users in Korea had no notion of the networking term "packet", which was used to quantify and bill for the service. Packets and their respective volume have little resemblance to what users experience on their devices based on these packets.

5.2 Focus groups

Focus groups are organised moderated discussions within a selected group of individuals to elicit their views, attitudes, feelings, beliefs, ideas, and judgments on a specific topic. Freely discussing and opportunistically exploring each other's views helps participants to reflect on their own opinions and the contexts that frame their presuppositions (Lunt & Livingstone 1996). '*Focus groups are particularly useful when there are power differences between the participants* [of a focus group] *and decision-makers or professionals* [that are interested in the participants opinions]*, when the everyday use of language and culture of particular groups is of interest, and when a researcher needs to explore the degree of consensus or divergence of opinion on a given topic'* (Morgan & Kreuger 1993). However, in a group setting, participants influence each other. It can thus be difficult to separate an individual's view from that of the group (Gibbs 2004). To overcome this problem, an adequate number of focus groups have to be conducted.

In order to get a first approximation of the key concerns of prospective users of mobile multimedia services, I conducted a series of 14 focus groups in London, UK. I supervised seven more focus groups led by a native French speaker following a set of guidelines (Knoche & Sasse 2004) at Alcatel Space (a MAESTRO partner) in France. Based on the same guidelines four groups were conducted by staff at Space Hellas in Athens, Greece. The results of the focus groups of all three countries are discussed in more detail in (Knoche & McCarthy 2004). The results presented here are based on the UK focus groups led and analysed by the author.

5.2.1 Participants

A total of 65 people (17 male, 48 female) participated in the 15 focus groups conducted in the UK, 90% of who were students at an average age of 25. Approximately 40% of these were Britons. The remainder came from Western and Eastern Europe, the US, Africa and Asia. (70%) had pay-as-you-go (non-contract) phones. The participants were recruited by means of a mailing list – the psychology subject pool at UCL - that people who interested in taking part in research subscribe to. The invitation targeted people that were interested in mobile TV services.

5.2.2 Procedure

The focus groups were structured according to the following stages:

1. welcome and scope of the occasion,
2. round-robin introduction of the participants,
3. general questioning on the broad topic,
4. specific questioning on suggested services and
5. summary and conclusion with a final individual voting on the proposed services.

To familiarize participants with one another and the subject at hand, participants were initially asked to introduce themselves (including their age and profession) and talk about their current use of mobile phones. The rest of the focus group session was more free-flowing, and only steered by the moderator with the following guiding questions (not necessarily in that order):

1. How do you currently use your mobile? What do you like what do you hate about it?
2. Have you ever switched your provider and if so, why?
3. How do you spend your commuting time, if you use your phone, how?
4. What are the current imperfections of the mobile lifestyle?
5. Would you choose a competing transport service, for example a bus operator if they provided a service like mobile TV?

Before these questions were discussed, I introduced the idea of mobile services involving television content by showing a sample clip on a PDA (HP Ipaq H2210, 64k colours, resolution 320x240, 116ppi) to each participant. This demonstration seemed necessary after the participants of the first two focus groups had been very sceptical about the attractiveness of watching TV content on a small screen. The scenario employed used the context of “dead time”, i.e., being in a state of limbo or waiting as experienced in commuting situations, waiting rooms, bus stops, lounges, *etc*. The participants were asked next whether they would use the following services and content types, visually supported by exemplifying pictures (see Appendix A 1). These services had been selected based on a brainstorming session with researchers:

1. live footage (e.g., sports events),
2. news and weather,
3. disaster relief information (an alert service that might turn on the phone),
4. music clips (MTV) - enhanced with a skip button to advance through unwanted tracks,
5. PVR functionality (that would allow one to record favourite programs – basically any content),
6. dating (that would allow for browsing multimedia content of prospective partners),
7. language courses (having the lecture's material on the phone).

To obtain some quantitative data, a ranking of the alternative services (similar to the nominal group technique (Delbecq & Van de Ven 1971); (Delbecq *et al.* 1975)) was obtained from individual participants at the end of each session.

5.2.3 Analysis of the focus group transcripts

After the transcription of the audio-recorded focus groups sessions my coding of the transcripts was informed by *grounded theory* (Glaser & Strauss 1999). I used key words, which described key concerns or qualities that emerged during the coding process. The individual codes were then grouped into broader categories and subsequent frequency counts of the codes within each category were used to weight the categories in importance.

The majority considered having a mobile phone obligatory, and usage did not necessarily imply being on the move. *Staying in touch* was their prime concern, which was challenged by *imperfect coverage*, *short battery life* and, indirectly, *cost*. The majority of participants chose their provider according to their social circle's preferences in order to control and minimise costs. *Stored content*, such as private SMS and contact information, was valuable to participants and they stated an interest in backup facilities, e.g. centrally with the provider. The ubiquitous availability model is marked by monetary and user costs. User costs on this level include charging the phone and carrying it around, as well as the task of remembering to do both. Watching sample video clips on an IPAQ (with a resolution of 240x320 pixels) made the generally sceptical participants more open to the idea of following video content on a phone. The moderator also suggested alternative viewing options like head mounted displays, projection techniques, and plugging into external displays.

1. Live footage

When asked what kind of content they would like to follow live, the majority said football. News was second for live footage, followed by other sports. However, for most people live content was something that they would rather experience in a group.

2. News & weather

Overall participants were most interested in news content (see Figure 18 for the total ranking distribution). Its timeliness, brevity and piecemeal-like character matched well with envisioned usage, e.g. while commuting, and the desire to be up-to-date. The benefit of having access to this content from abroad was also attractive.

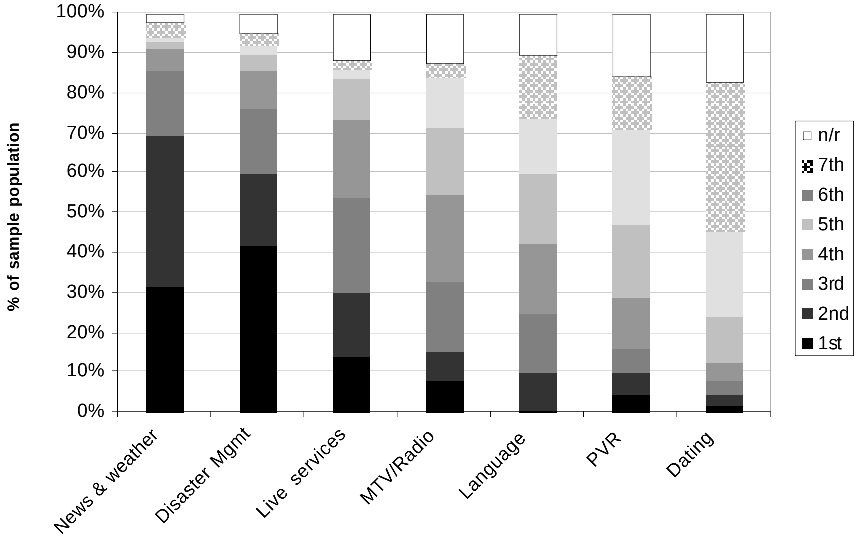


Figure 18: Total ranking distribution over all countries per service

3. Disaster management

A localised warning system used as a means to coordinate people around or away from disaster areas was also highly valued. However, the participants were not only worried about being under surveillance but also about who would be authorized to send warnings and whether the system could be hijacked or jammed. Another concern was the frequency of alerts with its effects on desensitisation and anxiety. In the absence of major disasters, they would like to be alerted of traffic irregularities. From a moral and democratic standpoint it was understood that charging for the disaster management service was inappropriate.

4. Music television

Music television enhanced by, e.g., a skip button, also has its fans but for the most part a convergence of radio and mp3-players with mobile phones was more appealing. The expectation was that content obtained through a service like this could be transferred to and from computers and that forwarding a song to a friend could not be prohibited.

5. PVR

A personal video recorder function that could record, e.g. television series, was most appealing to people with extensive commutes. Others found the concept of a mobile remote control for their home VCR attractive. In this context people were worried about a complex user interface.

6. Dating

A service that allows users to browse through multimedia files of prospective partners drew the weakest response, despite participants' claims that they knew many who would be interested in it. Many students asserted that their parents would consider it safer and easier to control than equivalent services on computers.

7. Language courses

This service was introduced as a fallback for missed language classes. Many of the group members, especially non-native English speakers, found this appealing, despite scepticism that they would want to expose themselves to demanding content while on the move. Discussions on this topic sparked ideas about having access to mobile dictionaries and audible translation services for words. Rather than use the service to make up for missed classes, participants expressed the desire to use the service as a general supplement to lectures, cooking programmes, and other how-to content.

5.2.4 Results

1. The participants wanted as large a screen as possible for viewing, but they did not want their phones to be too big.
2. They were worried about becoming too absorbed in what they are watching, and thus distracted from other tasks while being on the move, e.g. missing trains or stops. An easy way to set alarms or countdowns might help mobile users to not loose touch with the world around them.
3. They required a pause/mute facility to cope with likely interruptions. In the case of broadcast content, this requirement places demands on the device's storage capacity.
4. Volume control should be possible preferably without the need to access menus. The question whether a separate means to mute the volume and let the video play in the background will be necessary or might confuse users more in conjunction with the pause button, which pauses both audio and video has to be addressed by future research.
5. In terms of broadcast content news, sports and music were most interesting.

5.2.5 Discussion

Participants generally liked the idea of consuming multimodal content on their phone. However, it seems that for many, watching television on the phone was like learning to walk before you crawl – they were more concerned that coverage for standard calls and text messages fell short of their expectations. Listening to music or the radio while on the move was highly valued, and would require neither visual attention nor a significantly larger phone. Above all, battery consumption of multimedia services was a concern – participants were anxious that it might put their primary need in jeopardy - to stay in touch.

Participants' primary concern was cost and there was a strong interest in inexpensive multimedia content. Current price levels for multimedia services and inexpensive alternatives, such as newspapers, limited participants' enthusiasm. The size and weight of the phone required to use such services was a major concerns, especially for women. On the other hand, participants feared that small screen sizes and video quality would ruin visual details of the content. The results of the focus group are in line with the findings of the extensive mobile television study conducted in Finland (Södergård 2003).

5.3 Conclusions from initial user studies

From the background research and the initial studies, it became clear that the visual experience especially in terms of size will be an important concern for prospective users of mobile TV services. It is not clear what quality levels will be acceptable, and the users need to experience these first hand on mobile devices. A first lab study presented in the next chapter was designed around these parameters.

Chapter 6

An ant is over six feet tall when measured by its own foot-rule.
- Slovenian proverb

Study 1

Acceptability in relation to size

From the focus groups we have seen that image size and quality are amongst the most important factors in the QoE of mobile TV. Size, resolution and encoding bitrate affect perceived picture and video quality, but no study has systematically addressed these factors in conjunction on small screens. Previous studies (*cf.* Sec. 2.6.9) showed that perceived visual quality can depend on audio quality, too. The study presented in this Chapter investigated, which combinations of video encoding bitrate, image size and audio quality at a constant nominal angular resolution of 21ppd are necessary for different content types such that people find them acceptable for mobile TV services. The investigated content types included the three most popular content types from the focus groups - news, sports, music and I added animation content to the list. The value ranges of encoding bitrates for audio and video were informed by the results obtained during early studies in the MAESTRO project in which I was involved (McCarthy *et al.* 2004b). Previous research showed that people are willing to watch SDTV content at VRs between 11.7 (adults in Nathan & Anderson's study) and 5 (on 19 inch diagonal screen in Lund's study) with resulting angular resolution between 68ppd and 32ppd. The sizes selected in this study spanned a large range. Depending on the chosen viewing distance, the tested size range could result in a large overlap with living room VRs – albeit at a lower resolution due to both the display addressability and the content encoding.

6.1 Method

This study employed the acceptability rating method (*cf.* Sec. 3.5.1, p. 64) to assess video quality based on audio-visual stimuli. The aim was to evaluate the effects of varying image size, video and audio encoding bitrate on acceptability of the video quality of the service. I examined four different image sizes typical of current mobile phone screens (see Table 8) and represented roughly equal increments of pixel estate. The study did not control for viewing distance directly. As with normal mobile device use, participants were free to adjust the viewing distance and thereby VR to their individual preferences. Thus, prior to running the study, the viewing distances participants would adopt was unknown. However, a viewing distance of 40cm on the used device seemed reasonable and would have resulted in an angular resolution of approx. 31ppd. Westerink & Rouf had identified this (see Sec. 2.6.8) as the point at which the picture quality of slides presented on a projector cannot be increased by further reduction in angular size - by moving the picture further away - due to the effect of reduced picture angle. The average viewing distance observed in this study, however, was 27cm (see Sec. 6.7.1) and the values in Table 8 are based thereon.

Table 8: Image sizes used on PDA with estimated viewing distance of 27cm

Screen area in mm	Dimensions in and amount of pixels		VR	Ang Size	AR in ppd †
53 x 40	240 x 180	43,200	6.8	8.5 º	21
46 x 34.5	208 x 156	32,448	7.8	7.3 º	21
37 x 28	168 x 126	21,268	9.6	5.9 º	21
26.5 x 20	120 x 90	10,800	13.5	4.2 º	21

† The angular resolution represents a nominal value.

The encoding bitrate is an important factor because the effect of image size at a constant nominal angular content resolution might differ for the seven chosen encoding bitrates (*cf.* Sec. 4.6.3). At low encoding bitrates spatial resolution of the source video gets lost and not achieve the nominal angular resolution of 21ppd mentioned in Table 8. To which degree the encoding bitrate was sufficient to encode the spatial information for each encoding bitrate and content types is unknown (see also the discussion in Sec. 6.8). Encoding bitrate was manipulated in two different ways. Within each clip the bitrate allocated to video was reduced every 20 seconds by 32 kbps from a maximum of 224kbps down to 32kbps (see Table 9). The boundaries of the intervals were not pointed out to the participants; they were simply presented with a continuous clip that gradually decreased in quality over two minutes and twenty seconds. In addition to changing the video bitrate within a clip, two duplicate sets of clips were produced with different bitrates allocated to the audio channel, which were presented to different participants groups – a between-subjects condition (see Table 11). The *Low Audio* clips coded the audio channels at 16kbps Windows Media Audio (WMA) V9 whereas the *High Audio* clips were coded at 32 kbps. Theses values were selected based on results obtained from a study in the MAESTRO project (McCarthy *et al.* 2004b), in which participants' had rated the acceptability of audio when presented along with its video counterpart. The acceptability of audio at 32bps compared to 16kbps had dropped from 95% to 80%. Although in this study the primary task of participants was to rate the acceptability of the video quality, the aim of this between-subjects factor was to examine whether low audio quality would bias participants' perception of the video quality as indicated by e.g. (Neumann *et al.* 1991) and (Reeves *et al.* 1993) or rather follow the additive and multiplicative multimedia models described in Sec. 3.8. Finally, the participants were video recorded while they watched the clips to make obtain viewing distance estimates under the different conditions.

Table 9: Encoding bitrates for segments

Int.	*Time (secs)*	*Encoding bitrate video*	*Encoding bitrate audio*
1	1-20	224 kbps	16 / 32 kbps
2	21-40	192 kbps	16 / 32 kbps
3	41-60	160 kbps	16 / 32 kbps
4	61-80	128 kbps	16 / 32 kbps
5	81-100	96 kbps	16 / 32 kbps
6	101-120	64 kbps	16 / 32 kbps
7	121-140	32 kbps	16 / 32 kbps

6.2 Material

The study used content recorded directly from TV or DVD without any special editing steps. Clips of this type have been successfully used to examine quality tradeoffs for football coverage on mobile TV (McCarthy *et al.* 2004a). The length of the clips was based on Södergård's findings (see Sec. 5.2) that mentioned watching time between two and five minutes. News was the most demanded content class by all participant groups in different studies. Other content of interest were *sports highlights* and *music* *videos*. As an additional category stop-frame animation (claymation) was included as a category. Animation can be very bandwidth efficient. In total, I prepared four clips for each content type for a total of 16 source clips. Table 10 provides a summary of these clips.

The video clips were prepared as follows: Footage was recorded from digital terrestrial TV (BBC24 News) and from DVDs (2002 Fifa World Cup football, Creature Comforts animation, Michael Gondry music videos). All extracted clips were chosen such that after 2:20min (or shortly thereafter), a story line would end. I used *Virtualdub* to deinterlace and segment these source clips into seven 20 second long clips at the different sizes at 12.5fps. Windows Media Encoder (WME) encoded these segments with the Microsoft Windows Media Video (WMV) V8 codec with the different bitrates for the different segments as shown in Table 11. Each group of seven WMV segment files were then converted and concatenated to one AVI file using TMPGEnc Express. Finally, these files were encoded using WME again to alter the audio encoding to either 32 or 16kpbs using WMA V9 codec. The video was encoded at a higher bitrate than the maximum of the first WME encoding in order to prevent significant alterations in the video quality in any of the segments.

Table 10: Used content types overview

Clip	*Content Type*	*Description*	*Source*
N1-N4	News	BBC News 24 Headlines	DVB-T (freeview)
S1-S4	Sport	Football World Cup 2002: Goal Highlights	DVD
M1-M4	Music	Clips directed by M. Gondry	DVD
A1-A4	Animation	Clips from “Creature Comforts”	DVD

6.3 Experimental design

I chose the experimental design based on my discussions with John McCarthy. As shown in Table 11 there were four different groups, each comprising 32 participants. Each group was presented 16 clips in total in groups of four clips at each of the four image sizes. The groups differed in whether they experienced *Increasing* or *Decreasing* image sizes and whether the audio quality was *High* or *Low*. Within each group, a Latin squares design with four variations controlled for content such that the different content clips (e.g. N1-N4) were tested at each of the different image sizes across participants. The dependent variable was *Video Acceptability*. The independent variables were *Image Size*, *Content Types*, *Video Bitrate* and *Audio Bitrate.* Control variables were *Size Order*, *Gender*, and *Corrected Vision*. The variable *Corrected Vision* coded whether participants considered themselves to have normal vision or whether they wore contact lenses or glasses.

6.4 Equipment

An iPAQ 2210 with a 400Mhz X-scale processor, 64MB of RAM and a 512MB SD card presented the clips. The screen was a transflective TFT display with 64k colours and a resolution of 240x320 (116ppi). A set of Sony MDR-Q66LW headphones delivered the audio. A customized application was programmed in C# using the Odyssey CFCOM software (2003) to embed the Windows Media Player. It presented the clips along with a volume control and two response buttons to indicate acceptable and unacceptable quality. The program recorded at what time in which clip a participant clicked acceptable or unacceptable. A screen shot of the application is shown in Figure 19.

I undertook initial research to identify the different possible platforms for the development of an experimental application that would allow the play-back of video clips according to individual play-lists for the different participants and record their replies with time stamps. The two major choices that were available at the time were the *helix* player on the Symbian platform, and the windows media player on the MS .net platform, both of which offered SDKs. A middleware component called *odyssey,* which made the inclusion of the windows media player into .net applications easier, was provided free of charge for the research by the company odyssey.com. Two available iPAQ 2210 running Pocket PC 2003 were programmed with Microsoft's .net environment. Based on these parts Dimitrios Miras developed a first demo version of a mobile TV that would allow to play clips (McCarthy *et al.* 2004b). I then extended this version and developed the application for the subsequent experiments that would present individually pre-arranged play lists for each participant for the duration of the experiment and recorded their ratings with time stamps.

Table 11: Experimental design

	Audio	*Size*	*Dim.*	*Content Clip*			
A (32)	32 kbps	Dec.	240x180	N1	S1	M1	A1
			208x156	N2	S2	M2	A2
			168x126	N3	S3	M3	A3
			120x90	N4	S4	M4	A4
B (32)	32 kbps	Inc.	120x90	N1	S1	M1	A1
			168x126	N2	S2	M2	A2
			208x156	N3	S3	M3	A3
			240x180	N4	S4	M4	A4
C (32)	16 kbps	Dec.	240x180	N1	S1	M1	A1
			208x156	N2	S2	M2	A2
			168x126	N3	S3	M3	A3
			120x90	N4	S4	M4	A4
D (32)	16 kbps	Inc.	120x90	N1	S1	M1	A1
			168x126	N2	S2	M2	A2
			208x156	N3	S3	M3	A3
			240x180	N4	S4	M4	A4

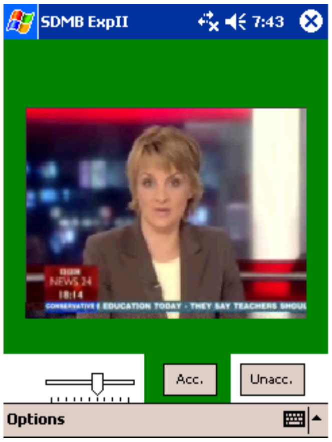


Figure 19: Application with volume control (lower left) and 'Acc.' and 'Unacc.' Buttons.

6.5 Procedure

The participants were told that a technology consortium was investigating ways to deliver TV content to mobile devices, and that they wanted to find out the minimum acceptable quality for watching different types of content. The instructions stated: *"If you are watching the coverage and you find that the quality becomes unacceptable at any time, please click the button labelled 'Unacc'. When you continue watching the clips and you find that the quality has become acceptable again then please click the button labelled 'Acc'.* Once it was clear that they understood the instructions, participants were provided with headphones and an iPAQ and given a short time to practice pressing the buttons on the display. When they were ready the experiment began and the participants watched 16 clips in succession.

During the session the participants' interactions with the devices were video recorded. The video was used to measure viewing distance. The participants' ratings, i.e. the taps on the 'Unacc.' and 'Acc.' buttons were recorded on the device. At the end of the video rating session, an interview followed to find out what aspects of the video quality they found unacceptable for the different types of content. The interview questions can be found in the Appendix (Sec. A 2, p.195).

6.6 Participants

Most of the 128 paid participants (83 women and 45 men) were university students. The age of the participants ranged from 18 to 67 with an average of 24 years. They came from a total of 26 different countries. English was the first language for 72 of the participants.

6.7 Results

Before analyzing the acceptability results, I conservatively labelled each 20 second interval of a clip *unacceptable* if the participant had given a rating of unacceptable at any point during that period or had ended the preceding period in the *unacceptable* state. The reported *acceptability* measures could therefore be interpreted as *the proportion of the participants that found a given quality level acceptable all of the time.* The resulting data was analysed using a binary logistic regression to test for main effects and interactions between the independent variables – *Image Size*, *Video Bitrate*, *Content Type* and *Audio Bitrate*. Control variables *Gender*, *Corrected Vision* and *Size Order* were also included in this analysis. Post-hoc within-subjects tests were performed using non-parametric Friedman and Wilcoxon tests.

The regression revealed significant effects on all of the control variables. *Gender* was a significant predictor of acceptability with women being less likely to rate a clip as unacceptable than men [$\chi^2(1)=12.6$, $p < 0.001$]. Participants wearing glasses or contact lenses (*Corrected Vision*) were less likely to rate a clip as unacceptable than those with normal vision [$\chi^2(1)=54.8$, $p < 0.001$]. Those participants who started with large image sizes that got smaller were generally more likely to provide 'unacceptable' ratings than those who saw clips increasing in image size [$\chi^2(1)=120.7$, $p < 0.001$].

6.7.1 Viewing distance

To see whether people compensated for the smaller size videos by holding it closer I scrubbed visually through the observational videos. I estimated that rough trends of pulling the screen closer in the decreasing and further away in the increasing size group should have been visible. This approach did not reveal any apparent trend and so I resorted to measuring the distance between the device and the eyes of the participant by using a ruler on the screen that was depicting the observational video. I sampled the video recordings of each participant randomly at the beginning and towards the end of the experiment to obtain these viewing distance estimates.

It turned out that almost all participants held the mobile device at a relatively fixed distance throughout the study. The average viewing distance was about 27cm and the range roughly between 15cm and 45cm. Since the participants were sitting on chairs, they either used their knees or the arm rest to support the hand holding the iPAQ, or rested the elbow of the arm holding the device on one leg. For both increasing and decreasing image size groups, there was no significant difference in the distance at which the iPAQ was held at the start or end of the study. Of those that frequently changed viewing distance throughout the study, this seemed to be more related to adopting a more comfortable posture while holding the device. Based on this average people watched content at best at a nominal angular resolution of 21ppd (*cf.* Table 11) but low *Video Bitrates* clips might have yielded lower spatial resolution (*cf.* Sec. 4.6.3). To ease comparison with previous research I will report the different sizes in the study in terms of their viewing ratio based on the average viewing distance across participants.

6.7.2 Size, video bitrate and content type

The logistic regression showed an expected significant effect of *Video Bitrate* on acceptability ratings [$\chi^2(6)=1186$, $p<0.001$]. However, there was also an interaction between *Video Bitrate* and *Image Size* [$\chi^2(18)=165$, $p<0.001$]. This interaction is illustrated in Figure 20 (left) for the two highest and lowest encoding bitrates. Averaged across content types, acceptability declined with decreasing image size at higher bandwidths. At the lowest bandwidth, there appeared to be a slight increase in acceptability. However, a post-hoc comparison revealed no difference between acceptability of the four image sizes at the lowest bandwidth [$\chi^2(3)=3.47$, $p=0.324$] indicating that there were no quality gains from reducing the image size in the parameter space tested in this study as outlined in Sec. 4.6.3 (page 89). Both *Image Size* and *Content* were significant predictors of acceptability, [$\chi^2(3)=446$, $p<0.001$; $\chi^2(3)=1056$, $p<0.001$] as was the interaction between *Image size* and *Content type* [$\chi^2(9)=136$, $p<0.001$] as shown in Figure 20 (right). The different content types have very different levels of acceptability. Not surprisingly, the low motion animation clips received the best ratings – for this type of content there was no significant difference in acceptability as image size was reduced from 240x180 to 168x126 [$\chi^2(2)=0.468$, n.s.], but at the smallest image size acceptability dropped off sharply [$Z=-6.49$, $p<0.001$]. For News content the *acceptability* significantly increased as the image size was reduced from 240x180 to 208x156 [$Z=-2.11$, $p<0.05$], after which point there was a steady decline in acceptability with decreasing image size. Thus, for News, I found some evidence that bandwidth savings might increase perceived quality. The curve for Music videos was relatively flat, and there was no significant difference in acceptability across the four image sizes [$\chi^2(3)=6.1$, n.s.]. Finally, Sports coverage showed the lowest levels of acceptability. There was no significant difference in acceptability between the two largest image sizes, but at image sizes smaller than 208x156 acceptability significantly declined [$\chi^2(2)=25.9$, $p<0.001$].

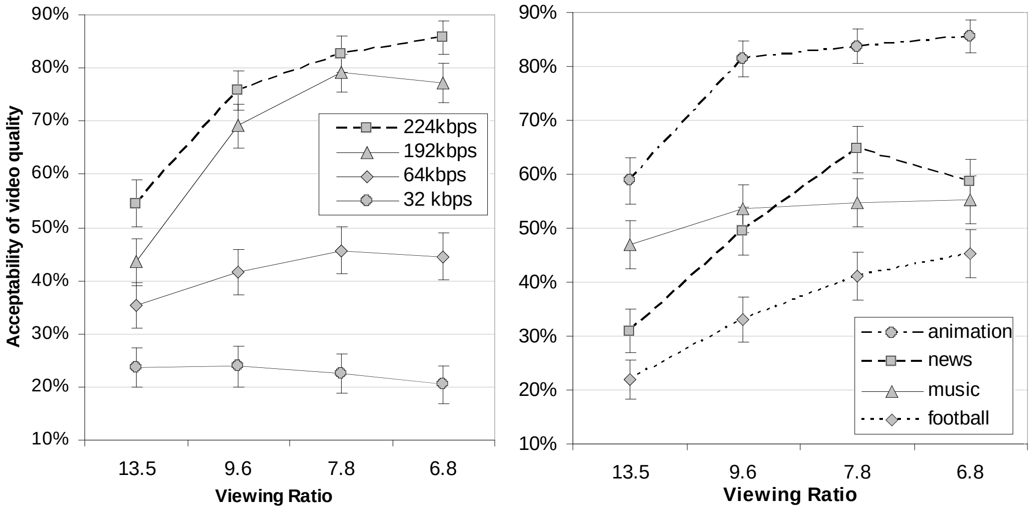


Figure 20: Acceptability of viewing ratios by encoding bitrate (left) **and content type** (right)

To illustrate these effects in more detail, the next sections present the content types separately at each of the seven video bitrates and the qualitative comments participants provided about why they found them unacceptable (see Figure 21 for an overview).

6.7.3 News

With News, the largest image size did not receive the highest acceptability ratings. Slightly smaller depictions 208x156 were more acceptable than 240x180. The effect was present at all video bitrates apart from 32 and 64kbps. Acceptability dropped considerably between 168x126 and 120x90. At 32 kbps no differences in image size were observable. When asked why they rated the News as unacceptable, participants mentioned a number of factors. Across all 128 participants, a total of 290 comments related to the unacceptability of News coverage. I coded these comments and grouped them into problem types. Figure 21 depicts a summary of these problems and their frequency. Of all comments, 34% related to the problem type *text detail*: the legibility of the news ticker, the headline text, the clock, the logo, or the captions for the people being interviewed by the newscaster. Other problems people commented on were *facial details* and expressions, the switch from anchor person to field reports (*shot types*), poor *audio fidelity* and a loss of *general detail*.

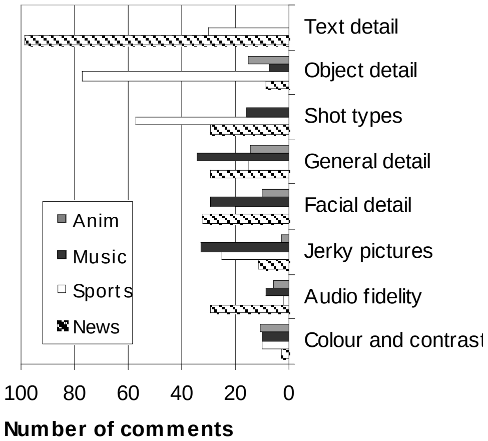


Figure 21: Reasons for unacceptable quality

6.7.4 Sports

With Football clips, acceptability increased with both *Image size* and *Video Bitrate.* However, even at the largest image size (240x180) and highest bitrate (224kbps) around 30% of participants found the quality to be unacceptable (see , right).

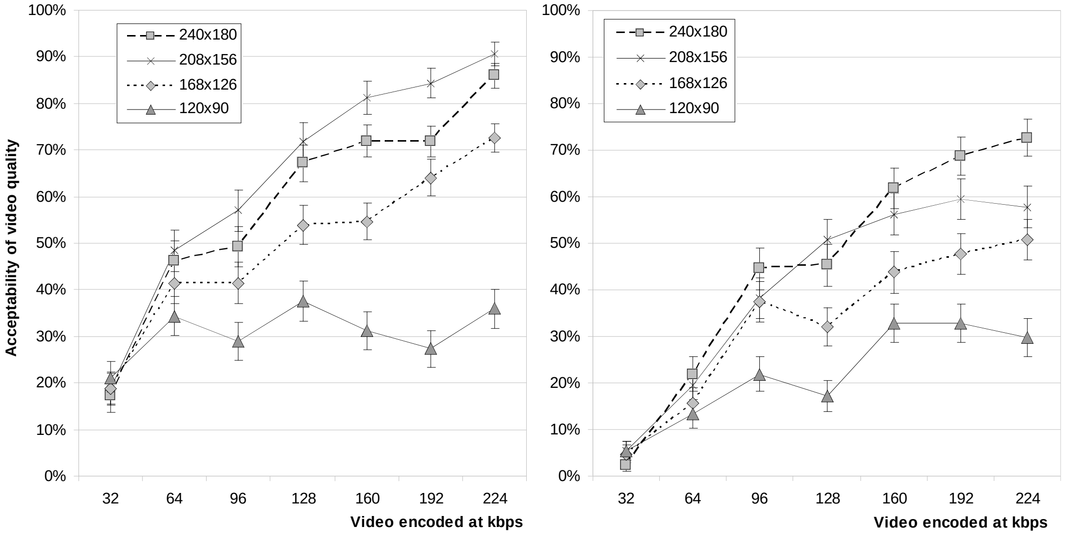


Figure 22: Acceptability of video quality of News (left) **and Football** (right)

The participants made 248 comments on the unacceptability of Football clips. The main problem was identifying *object detail*. In particular, participants reported problems seeing the ball and identifying players. The second most common complaints were about *shot types* - specifically XLS of the entire pitch, - which people found difficult to watch. In XLS it was harder to see the ball and identify players. Other problems included the inability to read *text detail* about teams and scores, the *jerkiness* of pictures and the inability to see *facial detail* clearly (see Figure 21). The second largest problem was insufficient object detail especially for the sports content. This might be especially a problem in conjunction with the next most frequent complaint about *shot types*. Extreme long shots are frequently used in sports coverage.

6.7.5 Music

With Music clips the effects of image size were less pronounced, but there was a clear interaction between *Image size* and *Video Bitrate*. At the lowest bitrate, the smallest images were rated as the most acceptable, but at the highest bitrate they were the least acceptable. Again this is evidence that higher encoding bitrates per pixel can improve the acceptability video quality – albeit at very low absolute values in this case. For Music clips, there were fewer comments on why quality was unacceptable. Of the 172 comments 34% mentioned *general detail* – such as blurriness and fuzziness - 33% related to the *smoothness of the frame rate.* Although the music clips were generally quite dynamic it was still interesting that the proportion of comments relating to frame rate (*'jerky pictures'*) was much higher for Music than Sports. Other major problems included the lack of *facial detail*, special effects and edits (*shot types*) and *colour and contrast* (see Figure 21).

6.7.6 Animation

With the Animation clips, a reduction in image size had little effect on acceptability apart from the smallest image size where there was a clear reduction in perceived video quality (See). Acceptability declined gradually with decreasing encoding bitrates with two exceptions. Acceptability dropped for all sizes when encoding bitrates went below 64kbps and a sudden drop occurred for the smallest size from 224 to 192kbps.

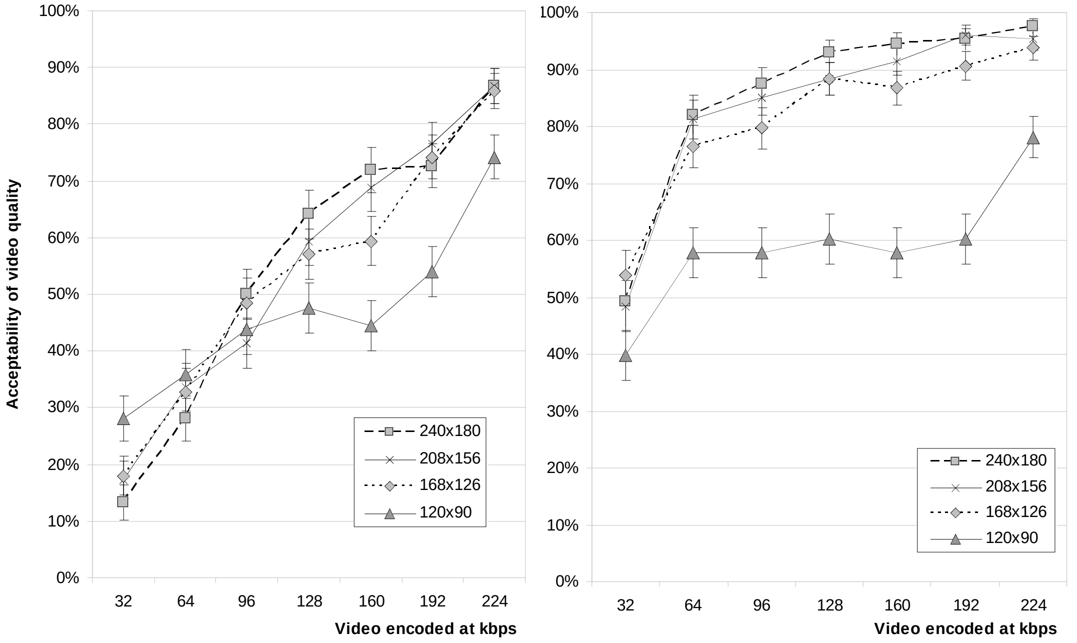


Figure 23: Acceptability of video quality of Music (left) Animation (right)

Animation yielded the fewest comments from participants in the qualitative interviews - only 64 in total, almost five times fewer than comments made about the News content. The most frequent complaint related to problems identifying the animal species in the animation when image size was very small. *General detail* was also mentioned and participants had problems when the image was very dark and the contrast was low (*Colour and Contrast*) *Facial detail* - such as the fidelity of the eyes and mouth - was also an issue as was the *audio fidelity,* which participants complained was *'echoic'*.

6.7.7 Qualitative feedback

In the qualitative interviews, participants made 147 comments that referred to experienced quality across all content types. The most frequent complaints were a general lack of detail, often referred to as a '*blurry*' or '*fuzzy*' display. A large number of comments specifically cited difficulty when the image size was small. In addition, almost 10% of comments complained about visual fatigue from watching small depictions – with problems such as '*It's tiring to watch*' and '*My eyes hurt*'. A further 8% complained about the effort involved when watching the very small screen with people complaining that they '*had to really concentrate to work out what was going on*'. As the viewing distance is relatively constant across different image sizes, this is probably not a problem of vergence, but of effort and fatigue from trying to decode information for viewing ratios as large as 13.5 at a low resolution.

Table 12: Problems across content

Problem	*% of general comments*
General detail	20%
Insufficient size	18%
Fatigue	10%
Effort	8%

6.7.8 Audio-visual interaction

Finally, there was a significant effect of *Audio Bitrate* in the logistic regression [$\chi^2(1)=62.8$, $p<0.001$]. As shown in Figure 24 at all video encoding bitrates the acceptability was rated higher when accompanied by the lower audio bitrate. This effect held across different image sizes and content types and was constant across the full range of bitrates, indicating that there is no interaction between audio and video quality in the parameter sub-space explored in this study.

6.7.9 Actual encoding bitrate

For all clip segments I calculated the actual encoding bitrate based on the file size and subtracted both the audio part and the overhead as stated by the encoder. The actual encoding bitrates of the clips did not always match the nominal encoding bitrates set in WME (see Figure 25). The target bitrates (grey diagonal in Figure 25) were both slightly exceeded by the larger size clips at smaller bitrates (above the grey diagonal) and not fully utilized by the smaller sizes especially at the higher bitrates (below the diagonal). Oddly, higher encoding bitrates did further raise the actual encoding bitrate closer to the target settings. I will consider the actual encoding bitrates in the discussion below.

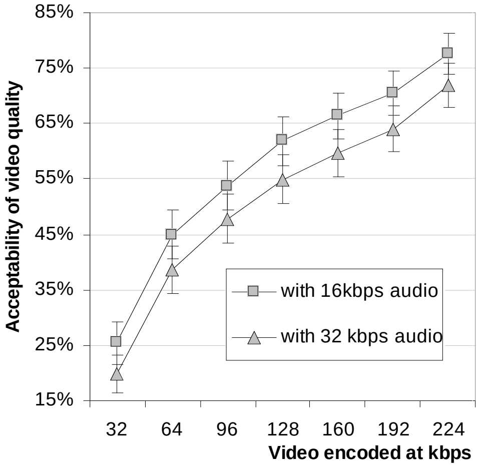


Figure 24: Video with different audio levels

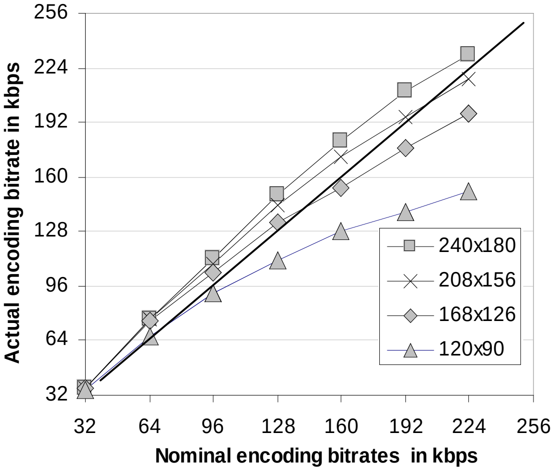


Figure 25: Actual *vs. n*ominal encoding bitrate averaged across content types

6.8 Discussion

I wanted to control for viewing distance to see whether people used it to compensate for small depictions of video and to arrive at preferred viewing ratios. Furthermore, the viewing distance estimates allowed for the computation of angular resolution at which the videos were watched.

The 27cm average viewing distance to the 115ppi device that I estimated based on the observation videos was smaller than the 35cm that Kato *et al.* (2005) observed on a higher resolution device (166ppi). It is also smaller than the 40cm suggested by Westerink & Roufs' optimal viewing distance with an angular resolution of 32ppd. In my study participants preferred to watch the content at 27cm, which resulted in viewing ratios between 13.5 and 6.8 and an angular resolution of 21ppd. A possible explanation might be the rating context. The participants could have held the device closer as they focussed on assessing the quality and provided feedback with the stylus on the screen. Clearly, people might chose a different viewing distance in real life while casually watching footage. This concern is beyond the scope of this study. Furthermore, I found no evidence for participants adjusting their viewing distance for the smaller image sizes. The smallest image size resulted in a VR of 13.5 higher than typical living room settings (*cf.* Figure 12, page 48). Again, this observation is limited to the context of rating the acceptability of video quality in the lab. The changing background colours depending on the acceptable or unacceptable state could potentially affect on the ratings by either providing a more or less favourable backdrop on the video. But since not a single one of the subjects mentioned or objected to the background in the debrief interviews this did not prompt further investigation.

Both quantitative and qualitative results indicated that the primary effect of reducing the image size was a loss of visual detail. Across content types, the effect of reducing image size was more pronounced at higher encoding bitrates. When the encoding bitrate was very low, there was little or no effect of reducing the image size, as visual detail was already poor. For all content types at 128kbps and above, there was a sharper reduction in acceptability when image size dropped from 168x126 (VR=9.3) to 120x90 (VR=13.5). The qualitative comments helped to identify the reasons. Of the eight most frequently cited problems, five relate to identifying or distinguishing detail – such as text, faces, players, animals and the ball. For News, Sports and Music, participants also identified particular shot types that caused difficulty. There were relatively few comments on frame rate, apart from Music clips, in which '*jerky'* frame motion seemed to be misaligned with the rhythm of the music and therefore disrupted the overall experience. Apart from News coverage, there was little evidence of any encoding bitrate savings or increases in perceived quality from reducing the image size. For News, the primary detail on which quality was deemed unacceptable was the ability to distinguish textual information – whether the news ticker, the clock, headline text or person names. The size of text should not be a concern since all but the smallest size resulted in VRs that are within the range of living room VRs. However, the resolution of the content (nominally between 240x180 and 120x90) and the angular resolution due to the display (about 21ppd) were much lower. For the two smallest sizes the ticker text was rendered on the device with only between 3 and 4 pixels in height, not enough to be legible. Furthermore, the angular sizes of text for 120x90 and 168x126 (9 and 13 arc minutes at the average VD of 27cm) were below the 15 arc minutes limit specified by Musgrave. It seems likely that the slight increase in perceived quality with a reduction in image size to 208x156 was caused by a perceived increase in the quality of the text. This could be attributed either to the higher encoding bitrate per pixel for the smaller size or due to unfortunate aliasing effects at the largest size. Study 2 looked into this in detail and compared transmitting text *separately* from the video and *inline* as in this study.

The acceptability of both Music and Animation at 120x90 dropped off significantly after the first 20 seconds at 224kbps. Since the actual encoding bitrate of the clips of this size was below the target (*cf.* Figure 25) the difference between 224kbps and 192kbps in terms of video quality should be marginal. Even for the resource demanding content type Football there was no significant difference in acceptability between these two values. It seems more likely that the difference between the two was an artefact of the experimental procedure namely the length of the quality intervals and the consecutive testing over 2:20 minutes or due to another hidden variable. The fact that it was too tiresome or unacceptable might have been apparent right within the first twenty seconds for News in which text legibility was poor and Football in which it was hard to identify players and the ball. For Music and Animation people might have realised this only after watching this size after 20 seconds. This highlights the potential value and need for assessing visual experiences with clips longer than 10 seconds often used in video quality studies following ITU recommendations.

Higher encoding bitrates result in better visual fidelity and are required to fully encode higher resolution content but how much additional encoding bitrate is required to support what video quality in higher resolution content is unknown. In this experiment I presented clips at their native resolution at seven different encoding bitrates and four sizes. Constant encoding bitrates applied to different resolution encoding settings as in this study confounds resolution and video quality. The resulting angular resolution in ppd of the larger sizes might have been lower than the smaller ones since they were constrained by the same encoding bitrate and might result e.g. in blocking artefacts. Last but not least the actual encoding bitrates of the videos differed from the nominal encoding (see Figure 25). The smaller the resolution and the larger the nominal encoding bitrate the larger was the deviation. Although the actual angular resolution of the video clips is unknown due to the influence of the spatio-temporal encoding I can arrive at better comparisons between the different sizes by considering the encoding bitrate available per pixel in the clips. Figure 26 depicts the acceptability values of the four sizes (excluding 224kbps) depending on the encoding rate (in bits per second) per pixel along with their logarithmic trend lines (fine). The bold curves are manually fitted trajectories for a given encoding bitrate allocated to different video resolutions. Moving from right to left on one of these curves means using a picture of higher nominal resolution (and in this study a larger picture). That a reduction of encoding bitrate resulted in the bold curves to approach plateau but not a decline suggested that the resulting actual resolution was not detrimental to the acceptability yet or was more than compensated by the increased size.

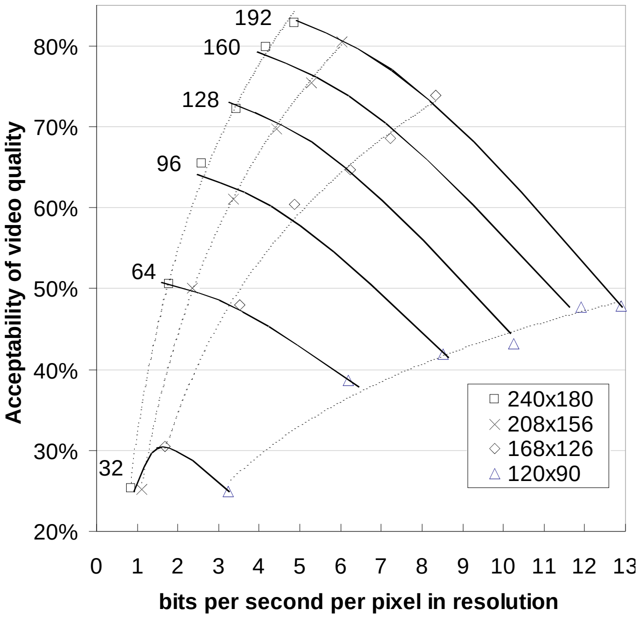


Figure 26: Trading-off actual encoding bitrate for size at a nominal 12.5fps

VQM can be used as an indicator for video quality (see Sec. 3.9). Researchers from TNO (Delft) provided VQM measurements of the video clips (de Koning *et al.* 2007). Figure 27 plots the averaged VQM scores (of 32, 64, 192 and 224kbps) on the x-, the viewing ratio of the four sizes on the y- and the averaged resulting acceptability values of the video clips on the z-axis. Acceptability depended both on viewing ratio and the visual quality of the clips. A value of 4.5 in VQM can - depending on the viewing ratio - result in acceptability values between 20% and 80%. Judging from the VQM scores the quality of the video is actually higher for the smaller sizes. However, the acceptability scores depend to a large degree on the size of the picture - further proof that the measure of acceptability as contextualized in this study incorporated the effects of both size and video quality. Few services might be interested in targeting customers with qualities that are only acceptable to less than 80%. From the evidence gathered so far there seems to be no beneficial trade-off in the observed parameter space at which lower resolution increased the video acceptability except for News.

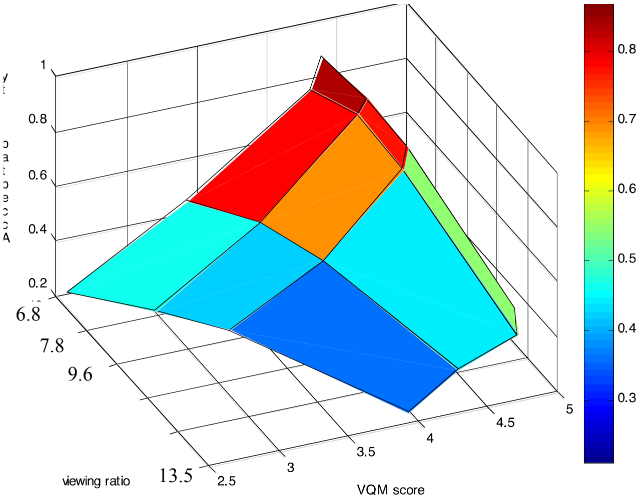


Figure 27: Contribution of size (in VR) and video quality (as measured by VQM) to acceptability

Overall, audio quality received few comments, with the exception of News in which audio generally carries the most important information. Participants were less likely to rate video quality as unacceptable when the audio quality was low (16kbps). This mirrors findings of Reeves *et al.* (1993) in which lower quality audio resulted in better assessments of video quality. One possible explanation of this effect is expected quality. Those given high audio quality might have higher expectations and are more easily disappointed or less forgiving with the visual counterpart. This is supported by previous findings by Bouch (2001) that showed the expectations had a significant effect on participants' ratings. However, as explained in Sec. 3.8 most multimedia models follow additive or multiplicative approaches.

One problem with the qualitative responses was the time of collection. People provided their feedback after having seen all video clips, all sizes and encoding bitrates. This made it hard to specifically attribute complaints to certain sizes and/or encoding bitrates. In study 6 (see Chapter 11) I solved this problem by collecting feedback for each quality setting individually while the participants were watching the clips.

From the results of this study it seems evident that people value watching mobile TV at sizes that result in similar viewing ratios as those in typical living room setups. Obviously the resolution of mobile TV as tried in this study leave a large gap to the limits of human spatial resolution perception. From the post-experiment qualitative feedback, the three most prominent problems relating to small size screens were identified.

1. Illegible text due to size, resolution or video quality. This will be addressed by the study presented in Chapter 7.
2. Shot types that do not scale well, Based on the data from this study I will extend the scope of the analysis by including the notion of shot types in Chapter 9.
3. Lack of detail. This led to the studies on zoom factors described in Chapter 10.
4. The viewing ratio from the chosen viewing distance resulted in a much lower angular resolution. The preference for size gives rise to the question as to how much more of the angular resolution can be relinquished to further increase acceptability through larger sizes. I explore how far videos can be scaled up or stretched in Chapter 11.

Study 6 in Chapter 11 is the one that most naturally follows the one presented in this chapter. However, in the course of my research I continued with study 2 presented in Chapter 7 in order to better understand the contribution of text on acceptability.

Chapter 7

Nothing in fine print is ever good news.
- Andy Rooney

Study 2

The influence of visual text quality on acceptability

News was identified as one of the most attractive mobile TV content types in the focus groups in Sec. 5.2. Study 1 showed that the acceptability of video quality depended largely on the type of content and that the legibility of text contributed to this. News below 168x126 pixels received low levels of acceptability. Many participants referred to text legibility as a criterion on which they had based their ratings. However, the analysis of the quantitative results could not answer the question whether the higher acceptability of the two larger clips was due to the legibility of text at larger sizes, or the higher resolution of the video. It was not clear whether 208x156 was more acceptable than 240x180 because of unfortunate aliasing at the larger size or because of the smaller encoding budget per pixel. The study also did not control for the participants' visual acuity. The degree to which imperfectly rendered text affects the overall perceived video quality is unknown. But it would be helpful for the planning of mobile TV services to know how the perception of different media influence each other, and how the bitrates allocated to the encoding video, including text, affect the service's acceptability. So far one cannot compare the cost and effort for preparing and sending text and video separately to the possible gains in acceptability. To address these trade-offs, I designed a study to assess image size requirements for mobile TV news. The aim of the study was to identify the impact of the visual quality of text on the overall acceptability. Because of the diversity in devices and bandwidth limitations the study tested different image sizes of mobile TV news and at a range of encoding bitrates. This sought to answer how much mobile TV news could gain in terms of perceived video quality from sending text separately through special formats, e.g. SMIL (W3C-Recommendation 1998) or QuickTime (Apple 2005) which can include text separately. To ensure the validity of the results, all tests were conducted on mobile devices. This chapter describes the method and the study on the effect of text quality at different sizes on video acceptability and presents the results. A discussion of the results follows in Sec. 7.7.2. The discussion in 12.5 includes overall recommendations for mobile TV news delivery.

Until now, only the study presented in Chapter 6 has shown that text quality might influence peoples' acceptability of mobile TV. However, it was not clear from the results of the study to what extent the text quality influenced the overall video quality perception because:

1. The study employed illegible text, fewer than five pixels in height.
2. Its participants were not tested for their visual acuity.

7.1 Method

The aim of this study was to evaluate the effects of text quality on the overall acceptability of video quality at different image sizes and encoding bitrates. A between-subjects design was used, in which one half of the participants saw news footage with *inline* text that degraded with the rest of the image. The other half experienced a simulated *separate* text delivery and saw the same footage with unimpaired text at high quality. The terms *inline* and *separate* text will be used to emphasize the implications for delivery and the corresponding conditions in this study will be denoted as *high* quality text and *degrading* text. The study employed the same method, sizes, encoding bitrates as used in study 1. The encoding bitrate of the video quality was gradually changed to find the cut-off point where the video quality became unacceptable.

This study used only one kind of device, which had a fixed resolution screen. In other words, the smaller size videos were represented by fewer pixels on the device. However, as with normal use the participants were able to freely adjust the viewing distance to the device such that the angular resolution (ppd) and VR could be changed according to their preferences. The VRs, angular resolution and text size indicated in Table 13 are based on the viewing distance of 27cm observed in study 1 on the same device in the same setting.

Table 13: Image sizes used on PDA based on an assumed VD of 27cm

Video area	*Pixels (P)*	*P/mm^2*	*VR*	*Ang. size*	*Ang. res.*	*text size**
53*mm* x 40*mm*	(240x180) 43,200	20	6.8	8.5 º	21 ppd	22.5’
46*mm* x 34.5*mm*	(208x156) 32,448	20	7.8	7.3 º	21 ppd	20.5’
37*mm* x 28*mm*	(168x126) 21,268	20	9.6	5.9 º	21 ppd	18.5’
26.5*mm* x *20mm*	(120x90) 10,800	20	13.5	4.2 º	21 ppd	17’

* Size (in arc minutes) of smallest text on screen, which was typically the ticker line

Although the participants’ primary task was to rate the acceptability of the video quality, the aim of this manipulation was to examine the effect of text quality on peoples’ perceptions of video acceptability.

7.2 Material

The four news clips that were used in the previous study, one of which included small text within the main window of the picture, were included such that comparisons between studies would be possible. Four additional news clips were recorded from the same digital TV channel in the UK (BBC24 news). These eight clips included a range of typical news coverage consisting of anchor person shots, stills, graphics, and field reports. Each clip lasted approximately 2:20 minutes.

The clips were cropped to 532x399 pixels from the original 720x576 to remove the letterboxing (the black bars that lie along the top and bottom of the screen) and to create a picture with a 4:3 aspect ratio. Figure 28 shows an example screen shot. The original ticker text had a size of 19 pixels in height. To control the influence of text quality I made the following alterations to the video material:

Step 1: In order to obtain content with an aspect ratio of 4:3 and an enlarged ticker the following parts of the picture were cropped off: 36 pixels from the left, 14 from the right and 14 from the bottom (the part below the ticker).

Step 2: The logo area in Figure 28, which contained both the logo and a clock, was overlaid by a bigger version containing only the logo.

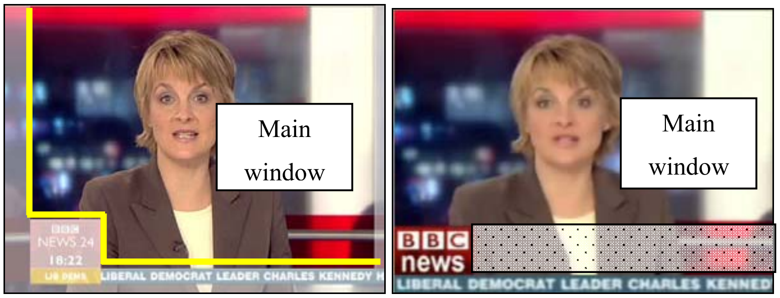

Main window
BBC NEWS 24 18:22
LIBERAL DEMOCRAT LEADER CHARLES KENNEDY H
Main window
BBC news
LIBERAL DEMOCRAT LEADER CHARLES KENNED

Figure 28: Video content before cropping and overlaying (left) **and after** (right)

Step 3: News headlines that were inserted in the area right of the logo (shown hatched in Figure 28, right) were overlaid with bigger font size versions that were still legible at the smallest clip size (120x90) with a size of 6 pixels equalling 17 arc minutes at a VD of 27cm.

Step 4: The area below the logo that featured a word to contextualize the ticker text was used to extend the space for the ticker as shown on the right in Figure 28.

Step 5: Varying lengths of the original ticker line were used such that in the final version the text ran across the whole horizontal length of the picture. Both the ticker and the main window were resized to their target size (see Table 13) using Virtualdub's bicubic resize. The height of the ticker line ranged from 9 to 12 pixels for the four sizes with the respective text height of the capitalized text ranging from approximately six to eight pixels. At a viewing distance of 27cm this resulted in a visual angle of the ticker text of 17 arc minutes for the smallest image size. The rest of the picture was slightly condensed in the vertical dimension to accommodate the ticker while meeting the target pixel estate laid out in Table 13. This allowed for comparisons of results with study 1 since the amount of presented information was approximately equivalent. In study 1 the size of the text in the ticker line ranged from 3 to 6 pixels rendering the text illegible for the sizes 120x90 and 168x126.

To ensure comparability of results with study 1, the audio that accompanied the videos was encoded at 32kbps in stereo (WMA V9). The video quality was manipulated in two ways. First, as in study 1 within each news clip the bitrate allocated to video was degraded every 20 seconds by 32 kbps from a maximum of 224kbps down to 32kbps (see Table 14). Second, in addition to changing video bitrate within a clip, two duplicate sets of clips were produced with different text qualities. Virtualdub segmented the source clips into seven 20 second-long clips at 12.5fps and for all sizes using a bicubic resize. These segments were encoded using WME, which used the WMV V8 codec with different bitrates for the segments as shown in Table 14.

Table 14: Encoding bitrates for video segments

Int.	*Time (secs)*	*Encoding bitrate video*	*Text Quality*
1	1-20	224 kbps	Separate/Inline
2	21-40	192 kbps	Separate/Inline
3	41-60	160 kbps	Separate/Inline
4	61-80	128 kbps	Separate/Inline
5	81-100	96 kbps	Separate/Inline
6	101-120	64 kbps	Separate/Inline
7	121-140	32 kbps	Separate/Inline

Each group of seven WMV segment files was then converted and concatenated to one AVI file using TMPGEncExpress. From these videos a second set was produced - with high text quality in which the ticker line, the BBC logo and text inserts above the ticker were replaced with the footage before the described degradation. Both the inline and the separately delivered text versions were then subjected to a final encoding using WME. The video was encoded at a much higher bitrate than the maximum of the first WME encoding in order to prevent significant alterations to the video quality. The difference between the footage at high and text quality are illustrated in Figure 29.

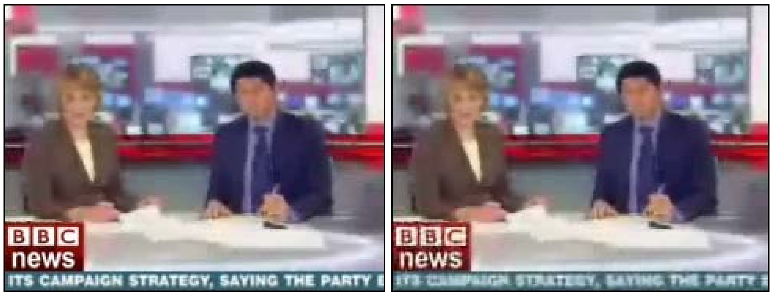

BBC
news
ITS CAMPAIGN STRATEGY, SAYING THE PARTY
BBC
news
ITS CAMPAIGN STRATEGY, SAYING THE PARTY

Figure 29: Comparison of material at 240x180 at 32kbps with separate (l.) and inline text (r.)

Two of the eight clips contained text in the main window that could be rendered illegible by smaller sizes. For better comparison with the results of study 1 the clips were included in the tested set and a control variable for them was included in the analysis. In the study presented in Chapter 6 the size of the text in the ticker line ranged from three (at 120x90) to six pixels (at 240x180), rendering the text illegible for the videos at sizes of 120x90 and 168x126.

7.3 Design

The experimental design followed the one used in Study 1. I ran four different groups, each comprising 16 participants (see Table 15). Each group viewed eight clips, in groups of two clips at each of the four image sizes. The groups differed in whether they experienced *increasing* or *decreasing* image sizes and whether the text quality of the ticker, the headline inserts, and the news logo was *Degrading* with the video quality or of constant *High Quality*. Within each group, eight variations controlled for content in a Latin squares design. This ensured that the different content clips n1-n8 were tested at each of the image sizes across participants. The dependent variable was Video Quality *Acceptability*. The independent variables were *Image Size*, *Video Encoding Bitrate* and *Text quality.* Control variables were *Size Order*, *Gender*, *Native English Speaker, Text in Content*, and *Standard Vision*. The control variable *Text in Content* identified the two aforementioned clips that contained small text in the main window. The variable *Standard Vision* coded whether participants had 100% visual acuity according to the administered two-eyed Snellen test (Bennett 1965).

Table 15: Experimental design

Gr.	*Text qual.*	*Res. Order*	*Image Size*	*Content Clip*	
A (16)	*seper.*	desc.	240x180	n1	n2
			208x156	n3	n4
			168x126	n5	n6
			120x90	n7	n8
B (16)	*seper.*	asc.	120x90	n1	n2
			168x126	n3	n4
			208x156	n5	n6
			240x180	n7	n8
C (16)	*Inline*	desc.	240x180	n1	n2
			208x156	n3	n4
			168x126	n5	n6
			120x90	n7	n8
D (16)	*Inline*	asc.	120x90	n1	n2
			168x126	n3	n4
			208x156	n5	n6
			240x180	n7	n8

7.4 Equipment

The test material was presented on an iPAQ 2210 with a 400Mhz X-scale processor, 64MB of RAM and a 512MB SD card. The screen was a transflective TFT display with 64k colours and a size of 240x320 (116ppi). The iPAQ was equipped with a set of Sony MDR-Q66LW headphones to deliver the audio. The same customized application as in study 1 was used. It presented the clips with a volume control and two response buttons to indicate acceptable and unacceptable quality, which were labelled 'Acc.' and 'Unacc.' (*cf.* Figure 19, page 99). While in the acceptable state the background colour around the video was green. This changed to red in the unacceptable state. The interface was designed to be in one of these two states at any given time.

7.5 Procedure

The participants completed a two-eyed Snellen test for 20/20 vision and an Ishihara test for colour-blindness prior to the start of the experiment. The participants were told that a technology consortium was investigating ways to deliver TV content to mobile devices, and that they wanted to find out the minimum acceptable video quality for watching news. The instructions were identical to study 1 and additionally stated explicitly that the participants could hold the PDA at any distance that was comfortable for them. Once it was clear that participants understood the instructions, they were provided with an iPAQ including headphones and given a moment to practice pressing the buttons on the display. When they were ready, participants watched eight clips in succession. Each clip started with the interface in the 'Acc.' state. The participants' ratings, i.e. the taps on the 'Unacc.' and 'Acc.' buttons were recorded on the device. At the end of the video rating session a semi-structured interview followed in which the participants were asked about the reasons for unacceptable video quality.

7.6 Participants

Most of the 64 paid participants (31 women and 33 men) were university students. The age of the participants ranged from 19 to 67 with a median of 25 years. The majority came from the UK (25) and China (18). English was the first language for 36 of the participants. Visual acuity was 100% for 48, 95% for seven, 85% for seven, and 80% or below for two of the participants. Two of the participants did not pass the colour-blindness test but their ratings appeared not to be outliers and were included in the analysis.

7.7 Results

Before analyzing the results, the segments' acceptability ratings were conservatively coded for each participant, i.e. each 20 second interval of a clip was marked as *unacceptable* if the video quality had been unacceptable at any point during that period or at the end of the preceding period. The resulting data was analysed using a binary logistic regression to test for main effects and interactions between the independent variables – *Image Size*, *Video Bitrate* and *Text Quality*. Control variables *Size Order, Gender*, *Visual Acuity, Native English Speaker* and *Text in main window* were also included. The regression revealed significant effects for the following control variables. *Size Order* was a predictor of acceptability [$\chi^2(1)=4.2$, $p<0.05$]. The participants who started with large image sizes that decreased during the experiment were generally more likely to rate the quality unacceptable than those who saw clips increasing in image size. *Gender* was a significant predictor of acceptability with men being less likely to rate a clip as unacceptable than women [$\chi^2(1)=45.5$, $p<0.001$]. The variable *Native English Speaker* was also a significant predictor for the experienced acceptability of the video quality [$\chi^2(1)=14.7$, $p<0.001$]. Native English speakers were less likely to rate the quality as unacceptable compared to the non-native speakers. Section 7.7.2 will address this in more detail. *Text in main window,* which was used to distinguish the three clips with text in the main window from the other clips, was also a significant predictor of acceptability [$\chi^2(1)=7.5$, $p=0.01$]. Videos without text in the main window were less likely to be rated unacceptable than those that did. The control variable *Visual Acuity* was not a significant predictor of acceptability [$\chi^2(1) =2.2$, n.s.].

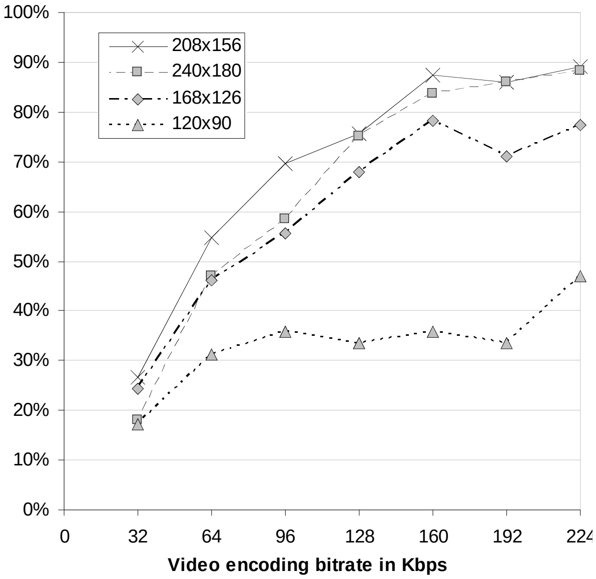


Figure 30: Acceptability of news at different encoding bitrates by size

As expected and in accordance with the results from Chapter 6, *Image Size* [$\chi^2(3)=270.7$, $p<0.001$] and *Video Bitrate* [$\chi^2(6)=414.6$, $p<0.001$] were significant predictors of acceptability. Figure 30 presents these results averaged across the two text delivery scenarios. Despite the legibility of the text in this study compared to study 1 the acceptability of video quality still dropped dramatically when image size was reduced to 120x90 pixels. This was similar to the results in study 1 but there the acceptability had already taken a sharp drop at 168x126.

7.7.1 The effects of text quality

Across all participants text quality was not a significant predictor of the acceptability of video quality [$\chi^2(1) =2.4$, n.s.]. This was due to the fact that the opposing ratings of the non-native and native speakers cancelled each other out. Post-hoc tests revealed an interaction between Text Quality and Native Speaker [$\chi^2(1)=40.1$, $p<0.001$] - illustrated in Figure 31. This effect came as a surprise. Native speakers who watched clips supported by separate text rated them higher in terms of acceptability than the non-native speakers. The non-native speakers rated video quality higher when video was accompanied by text that was presented inline and degraded with the video. The data was then partitioned and looked at separately for the

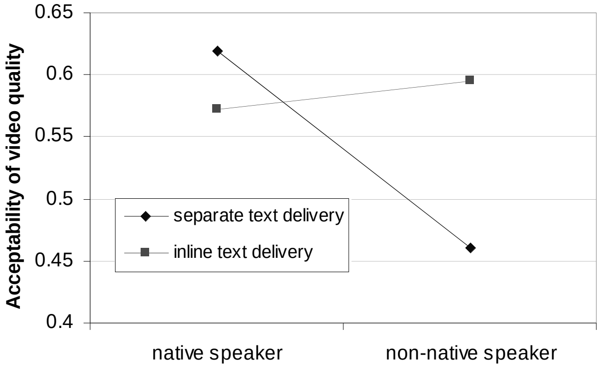


Figure 31: Interaction of Text Quality and lingual ability

two groups. Two non-parametric Mann-Whitney tests showed significant differences for *Text Quality* for both the native speakers [$Z=-2.1$, $p<0.05$] and the non-native speakers [$Z=-5.3$, $p<0.001$]. I ran the original binary logistic regression without the variable *Native English Speaker* on the partitioned data set of the native and the non-native speakers. Along with all the previously described variables *Text Quality* turned out to be a significant predictor of acceptability in the analysis of the native speakers [$\chi^2(1)=8.2$, $p<0.01$] and the non-native speakers [$\chi^2(1)=21.7$, $p<0.001$] - but as described above in opposing directions. Similarly, the control variable *Text in main window* was a significant predictor of acceptability [$\chi^2(1)=17.4$, $p<0.001$] for the native speakers but not for the non-native speakers [$\chi^2(1)=0.01$, n.s.]. Considering the impact of the non-native speakers the presentation of the following results will be limited to the 36 native speakers. Averaged across all encoding bitrates and sizes the acceptability of news content increased from 57% with inline text to 62% when presented with separately delivered text.

7.7.2 Qualitative results

Only a few participants - mainly from the degraded text quality groups - mentioned problems relating to text. The most common complaints about unacceptable quality related to low frame rates, audio-visual asynchrony, and loss of detail relating to eyes and lips, as well as the general inability to recognize people and objects or to identify who was speaking.

7.8 Discussion

This study investigated if and how the user experience of mobile TV news can benefit from high quality text provided by separate text delivery from the video and to disambiguate the results from study 1 in which 208x156 resolution content was more acceptable than 240x180 at constant encoding bitrates.

From the results of this and study 1 it can be concluded that text quality has a significant influence on the acceptability of video quality. Section 7.2 described the changes to the layout of the video clips. This resulted in the legibility of ticker text, the logo and inserted texts in terms of the visual angle at all sizes used in this study. This approach tried as much as possible to control for effects that illegible text might have on the acceptability. Now I can use the data to compare it to the results of study 1 and thereby measure the acceptability gains of these manipulations for the clips that had ticker text that was too small too read (i.e. at 168x126 and 120x90). To perform a fair comparison, I included only the ratings of native speakers for the four clips that had been used in both studies. Figure 32 shows a plot of the acceptability values averaged across the two smallest sizes and all encoding bitrates for both the three clips with no small text and the one clip that did have small text in the main window. When compared to the results obtained in Study 1 (in Chapter 6), the simple manipulations to ensure ticker legibility led to a substantial increase in acceptability. More than 20% more of participants (from

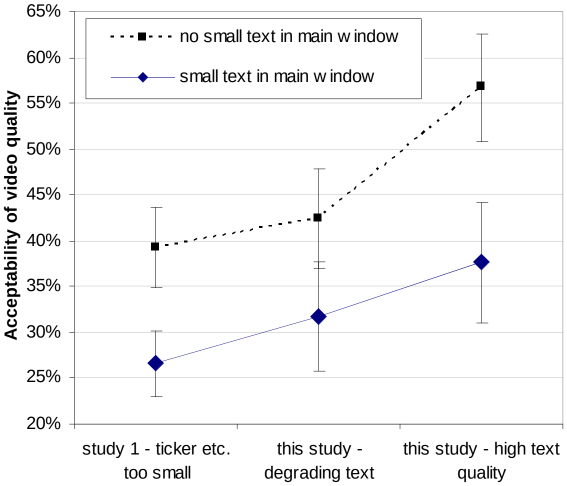


Figure 32: Acceptability gains of restructured layout and contribution of text to acceptability

37% to 45%) found the video quality acceptable when the ticker was large enough to read and degraded along with the video. When the ticker text did not degrade with the rest of the displayed video, native speakers' acceptability ratings increased by another 26% in relative terms compared to the clips with degrading ticker text. Finally, compared to the original study 1 complaints about text legibility had decreased in numbers tremendously.

Non-native speakers did not seem to include textual quality considerations into their video acceptability ratings because the *Text in main window* variable was not a significant predictor in their quality ratings. Native speakers rated video quality lower when text was shown in the main window and perceived an overall increase in quality when the text quality of the ticker text was high. Whereas the native speakers might appreciate better legibility and become more immersed in the content non-native speakers might read the ticker less and concentrate on the difference in quality between the main window and the overlaid area. The cognitive dissonance induced by the perceived quality mismatch of these areas could account for the differences in perceived quality. The high quality of the text could set the quality baseline for the rest of the picture higher. But there are other explanations for why native speakers would rate video quality higher than non-native speakers in general. One should note that the participants might have interpreted acceptable in different ways – for the native speakers the utility of the footage might have played a bigger role than for the non-native speakers that might have focussed on the visual quality. Non-native speakers could simply be less immersed in the content and therefore focus more on rating the quality. It could also be explained by non-native speakers' expectations of the video quality. Another explanation could be a more holistic rating of the video quality. For example, Chua *et al.* have shown that participants from the western hemisphere typically focus more on the foreground of videos and those from the eastern hemisphere make more balanced judgments (Chua *et al.* 2005). This could be easily tested by running another study with e.g. Chinese content including text and Chinese participants. For video quality studies this poses an interesting problem when text is part of the video clip.

The fact that non-native speakers rated the overall acceptability lower than native speakers gave rise to a re-analysis of the results of news content in study 1. It turned out that only the non-native speakers had found news content at 208x156 more acceptable than at 240x180. At 208x156 the text in height was 20.5 arc minutes and for 240x180 about 22.5 arc minutes. Among native speakers the two sizes were equally acceptable. Recall that the text sizes in study 1 were smaller than in this study.

In Figure 33, I have collated the acceptability results of the native speakers that saw the news clips with high text quality (left) and degrading text quality (right) each depending on the bits per second per pixel in resolution. I have included logarithmic trend lines (fine) for each size and bold ones for each encoding bitrate. The most obvious difference between these two conditions is that for encoding bitrates of 128kbps and higher the acceptability reaches a plateau (Figure 33, right) once size gets bigger than 208x156 with high quality text. But it declined when the text quality degraded with the video (Figure 33, right). I can therefore assume that the decline in text quality was responsible for the differences in acceptability and that the encoding bitrate was not high enough to sufficiently encode the text for 240x180 at the highest encoding bitrates (>96kpbs). In fact an encoding bitrate per pixel of 7 or more would have been required to render the ticker and other text at a sufficient quality as can be derived from Figure 33, right. For encoding bitrates of 96kbps and smaller the overall picture quality of 240x180 was lower than at 208x156 and text quality did not alter this.

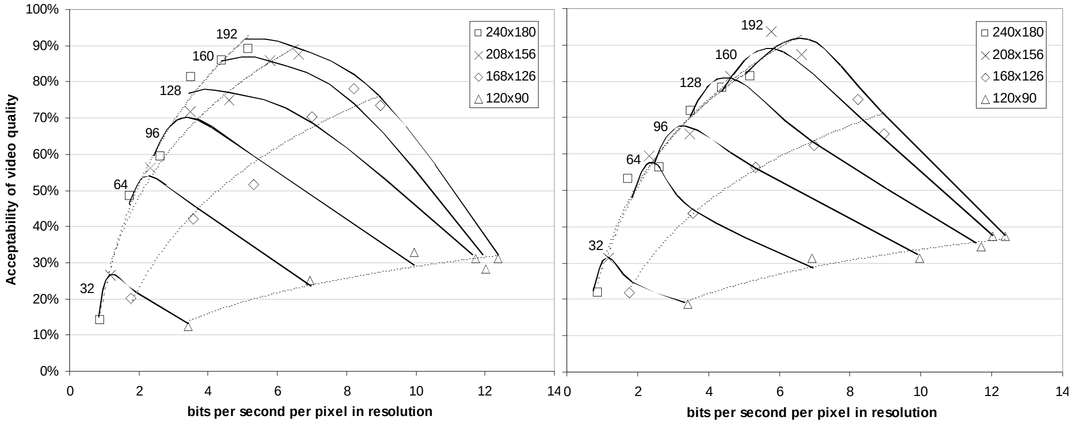


Figure 33: News video acceptability for native speakers -
with *separate* (left) and *inline* text delivery (right)

The study did not control for the possible effects of aliasing that might have been introduced from the bicubic resizing. But since the emphasis of the study was comparing the high and degrading quality versions of the same underlying text the influence of this was controlled for as far as comparisons within a given size were concerned.

7.9 Conclusion

Study 1 showed that native speakers found the largest size of news the most acceptable because of the legibility of text. In this study, text was 15 and 16 arc minutes in height and size therefore less of a possible concern. When text was delivered in-line native speakers found news at 208x156 most acceptable because of the better definition of text. The trade-off mentioned in Sec. 4.6.3 (p. 89) about increasing the encoding bitrate per pixel by reducing the overall resolution worked for small detailed text. In the next chapter I will present the results of a field-based study in which four of the news clips used in the current study were combined with the twelve sports, animation and music clips used in study 1. These clips were shown to people while travelling on the London underground network.

Chapter 8

Study 3
Watching on the train

The results of both Study 1 and 2 were obtained under lab conditions. Although this controls many variables such as lighting conditions, noise, external distractions it is not clear in how far the results would hold in a more realistic setting. To test the ecological validity of the results from study 1 and 2 I conducted a study with 32 participants that were watching the same material on the London underground.

8.1 Design

The experimental design followed the original lab study. I ran two groups: each group of 16 participants viewed 16 clips in groups of four, at each of the four sizes. The groups differed in whether they experienced *increasing* or *decreasing* image sizes. Within each group, I ran eight variations to control for content using a Latin squares design. This ensured that the different content clips were tested at each of the image sizes across participants.

8.2 Material

I re-used the animation, football and music clips from study 1 and included the 4 news clips from study 2, which had been shown in study 1 but with different ticker and logo text (see Sec. 7.2 for details). The video clips were encoded at four resolutions (240x180, 208x156, 168x126, 120x90). Within each clip, the bitrate allocated to video was gracefully degraded every 20 seconds in steps of 32 kbps from a maximum of 224kbps down to 32kbps. The boundaries of these intervals were not pointed out to the participants. They were told only that the quality would vary over time and were presented with 16 clips, each of which gradually decreased in quality. For all clips the audio was encoded at 32kbps in stereo (WMV V9). The details about the productions of video clips can be found in Chapter 6 (study1) and for the news clips in Chapter 7 (study 2).

8.3 Equipment

The test material was presented on an iPAQ 2210 with a 400Mhz X-scale processor, 64MB of RAM and a 512MB SD card. The screen was a 116ppi transflective TFT display with 64k colours and a resolution of 240x320. The iPAQ was equipped with a set of Sony MDR-Q66LW headphones to deliver the audio. I used the same interface as in study 1 and 2 to present the clips. The interface offered two buttons, which allowed the participants to switch back and forth between acceptable and unacceptable feedback with little effort.

8.4 Participants

32 paid participants (11 women and 21 men, age 20 to 65 with a median of 28 years) were university students. The majority came from the UK (20) and English was the first language for 28 of the participants. Visual acuity was 100% or higher for 24, 95% (6), 90% (1), and 85% for one participant.

8.5 Procedure

Before boarding the London Underground trains, participants were instructed by the experimenter, who accompanied them at all times. The participants were told that a technology consortium was investigating ways to deliver TV content to mobile devices, and that they wanted to find out the minimum acceptable video quality for watching mobile TV content. The instructions were identical to study 1 and 2 and included that the participants *"... can hold the PDA at any distance that is comfortable for you."* Each clip started with the interface in the 'Acc.' state. The participants watched eight clips on the outbound journey, and another eight clips on the return train. After the first eight clips the application stopped playing to make for a safe transfer. The train journeys included both underground and over-ground segments. Throughout the experiment on the trains the participants were video recorded and a debrief interview concluded the session about which aspects of the video quality they had found unacceptable. I also asked whether they had had any specific problems watching while riding on a train. The route was chosen according to availability but all the rides included both under and over ground segments.

8.6 Results

I combined the acceptability ratings from the train with data obtained in study 1 and 2. I included the acceptability ratings from the football, music and animation clips of the 64 participants that had experienced video accompanied with 32kbps audio in study 1. From study 2 I included the acceptability ratings of the four news clips (that were shown on the train) from the 32 participants who watched degrading text quality in the lab.

The combined results from the train and the two lab studies were then analyzed on a second by second basis using a binary logistic regression to test for main effects and interactions between the independent variables of the previous studies – *Image Size*, *Video Bitrate* and *Content Type* and *Context*. Context denoted whether the data was obtained in the lab or the field. Control variables *Gender, isNativeSpeaker* and *Size Order* were included in the analysis.

The regression revealed significant effects on all of the control variables.

1. Women found the video quality more acceptable than men [$\chi^2(1)=185.6$, $p<0.001$],
2. non-native speaker more than native speakers [$\chi^2(1)=185.6$, $p<0.001$] and
3. the people whose clips increased in size more than those whose clips decreased in size during the experiment [$\chi^2(1)=2615.1$, $p<0.001$].

As in the previous studies *Image Size* [$\chi^2(1)=5377.7$, $p<0.001$], *Video bitrate* [$\chi^2(1)=16.7$, $p<0.001$], *Content Type* [$\chi^2(1)=5377.7$, $p<0.001$] and the interaction of *Image Size* and *Video Bitrate* [$\chi^2(1)=2309.6$, p<0.001], were significant predictors of *acceptability*. Larger *image sizes* and higher *video bitrates* resulted in higher *acceptability*. But at low *video bitrates* the benefits of larger *image sizes* diminished.

It turned out that *Context* was a significant predictor of *acceptability* [$\chi^2(1)=502.1$, p<0.001] – the participants found the quality of the clips more acceptable on the trains than in the lab. This was not true across the board as the interaction of *Context* with *image size* was also significant predictor [$\chi^2(1)=2309.6$, p<0.001] of *acceptability.* For the smaller *image sizes* there was no significant difference but a non-parametric Kruskal-Wallis test [$\chi^2(1)=24.56$, p<0.001] showed that the participants on the train found the larger two sizes more acceptable than the participants in the lab. This finding is summarised in Figure 34 (left).

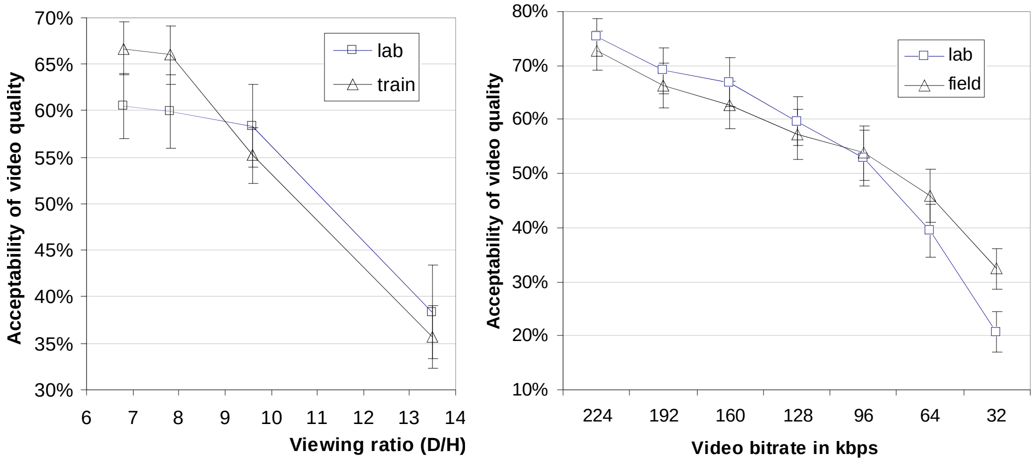


Figure 34: The interaction of *image size* (in VR based on *D*=27cm from study 1) and *context* (left) **and of *video bitrate* and *context*** (right)

The interaction of *Context* with *encoding bitrate* significant was another significant predictor [$\chi^2(1)=306.9$, p<0.001] of *acceptability.* At high video bitrates there was no difference between lab and field but for low video bitrates (<100kbps) the participants on the train found the video quality more acceptable than the participants in the lab (*cf.* Figure 34, right).

The biggest problems to an acceptable experience mentioned in the qualitative feedback were the unsteady motion and noise induced by the train and the stations and the reduced contrast due to sunlight shining on the screen. Holding the device closer and using a second hand to shield the sunlight helped partly but the people found it to be a tiring solution. Viewing in tunnels was deemed far superior but at the same time made the shortcomings in bright daylight more apparent. Nevertheless, people got immersed in the content and some expressed worries that they would miss their stop if they used mobile TV on public transport.

8.7 Discussion

The acceptability ratings for video quality in the lab were generally lower than those obtained on the train. This is in line with results of Jumisko-Pyykkö *et al.* (2008), whose participants rated the audio-visual quality of clips impaired by packet loss consistently higher in the three contexts in the field (bus, train station, cafe) compared to the lab. The difference was most pronounced at the lowest quality – the highest loss ratio (Jumisko-Pyykkö & Hannuksela 2008). I found the same to be true for low encoding bitrates, which were more acceptable on the train than in the lab. For service providers delivering video content in medium to high quality, my lab results provide conservative estimates of the levels of quality, which their customers will find acceptable when viewing on the move. In terms of the size requirements the story was different. My results showed that on the train, acceptability already declined once the viewing ratio was larger than 8 and the larger sizes (208x156 and 240x180, VR<=8) yielded a higher acceptability than in the lab. The acceptability of depictions smaller than 34.5mm (VR>8) resulted in equally reduced experiences both in the lab and on the train. Further research is required to find the reason behind this but the qualitative feedback I received points at viewing while in motion, under noise and reduced contrast due to sunlight as possible avenues to explore.

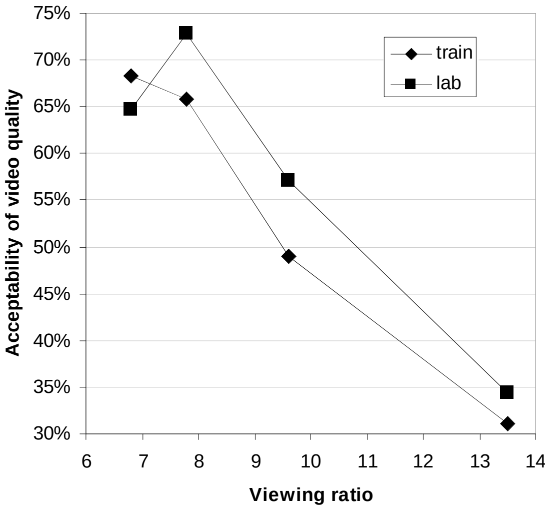


Figure 35: Acceptability of news on the train and in the lab judged by native speakers

News was a special case in study 1 in that smaller (208x156) depictions were more acceptable than larger (240x180). Study 2 showed that this was due to the fact that the encoding bitrate was not high enough to render the text sufficiently. Since size turned out to be more important in this study I will revisit the trade-off between size and visual quality. News was the only of the four content types that was overall less acceptable on the train than in the lab. Figure 35 depicts native speakers' acceptability ratings of news content accompanied by inline text on the train and in the lab (in study 2). Unlike in the lab the largest size (VR=6.8) with text of 22.5 arc minutes was the most acceptable on the train despite its reduced visual quality. This again emphasizes the added value of size in the field that I found overall in this study.

One of the limitations of this study is that I did not control for ambient lighting in my analysis. Lighting changed both with watching under and over ground and in the latter case further with sunny and overcast days. However, my finding about low encoding bitrates being more acceptable on the trains than in the lab should hold because the encoding bitrate segments were short, repeated many times for each participant and should have there been subjected to both high and low ambient lighting in the field equally. In terms of the contribution of size on the train in comparison to the lab the largest (240x180) and the smallest size (120x90) might have been shown underground for more time than the two sizes in between - the outbound route started and the inbound route ended underground. But since the acceptability scores of the largest and second larges size on the train did not differ significantly a large confounding effect of ambient light on at sufficient sizes seems unlikely. Although, as a standalone experiment this study would have benefited from either randomizing the size order or explicitly controlling for ambient lighting this would have made complicated the comparison with the results from study 1and 2, both of which used steadily increasing and decreasing sizes. Future research should further evaluate the effect of ambient lighting on the acceptability of the visual experience.

In the next chapter I re-analyze the data from study 1 based on shot types.

Chapter 9

Tragedy is a close-up; Comedy, a long shot.
- Buster Keaton

Study 4

The effect of size and resolution on shot types

The participants' most frequent complaint in study 1 presented in Chapter 6 was the unsatisfactory rendering of certain shot types on the small screens. Especially for football content, extreme long shots were rated unacceptable. Prompted by these complaints, the dataset was analyzed according to the shot types introduced in Sec. 0.

Current editing rules are not based on empirical research, but on rules of thumb or expert opinion. The following analysis investigated the acceptability of directly recorded TV or DVD material without any special editing steps to see how the different shot types would be affected by the different encoding settings - in particular size and resolution. This was compared to objective video quality measures obtained through PSNR.

9.1 Material

There are two possible caveats with the approach used here.

1. Due to the use of acceptability (*cf.* Sec. 3.5.1) on video clips with degrading video quality, the experimental design did not present all parts of each video clip at all encoding bitrates. Consequently, the average encoding bitrate at which shot types were encoded were not identical.
2. Video encoders compress e.g. low motion video clips better than clips that include a lot of motion. Some shot types might contain more motion on average than others and therefore look better after encoding in terms of visual quality, e.g. sharpness. Thus even if the shot types had been encoded at identical average encoding bitrates would have not guaranteed equal visual quality of the shot types after encoding.

To control for both differences in encoding bitrate and possible correlations between shot types and encoder performance, I used the objective quality measure PSNR (*cf.* Sec. 3.9) to obtain rough estimates of the content's visual quality.

There were two possible approaches to comparing the different size clips:

1. Compare each degraded clip with a down-sized version of the original clip.
2. Up-sample all degraded clips to a common bigger size, and compare it to the original clip at that size.

The problem with the first approach would be that it would not deliver comparable results between image sizes, since each file would be compared to a different original. Furthermore, the smaller size clips would yield a smaller error than the bigger size clips.

Therefore, all degraded clips were rescaled up to the size of the original clips, and Avisynth's built-in PSNR compare function computed the degradation of these encoded clips in comparison to their originals (Avisynth 2005). Since this approach compared up-scaled versions of the small size clips with the reference clip, one can expect that the smaller size clips will in general yield lower PSNR scores. For example, a clip with a size of 120x90 would be up-scaled by a factor of about four, which will result in higher peak signal-to-noise ratio than a clip up-scaled from 240x180 by a factor of two. The obtained PSNR scores were only used as indicators of visual quality between the shot types in clips of the same size. The PSNR values of the different shot types for the different content types are presented in the following section.

9.2 Results

The data were generated from the acceptability replies of the participants obtained in the study presented in Chapter 6 on a per second basis. For example, when a participant clicked unacceptable during the 35th second of a clip I marked all the previous seconds 1-34 acceptable and all the following seconds from 35 to 140 as unacceptable. I decided to exclude all ratings in the three seconds following a scene change to allow for participants' adjustment to the new picture. This consequently excluded shots that lasted fewer than three seconds. For the analysis of the resulting data the variable *Shot Type* was included as an independent and *Native Speaker* as a control variable. The latter variable denoted native English speakers. Both variables were in addition to those variables, which were analysed in Chapter 6. Analogously, the data were analysed using a binary logistic regression to test for main effects and interactions between the independent variables – *Image Size*, *Video Encoding Bitrate*, *Content Type, Shot Type* and *Audio Bitrate.* Control variables *Gender, Corrected Vision, Size Order* and *Native Speaker* were also included in this analysis. The variable *Corrected Vision* indicated whether participants had uncorrected vision or wore contact lenses or glasses.

The regression revealed significant effects of all the control and independent variables, as reported in Sec. 6.7. Non-native English speakers were less likely to rate the quality of a clip unacceptable than the native English speakers. The data from the non-native speakers were excluded and the regression repeated. All results presented from hereon are based on the 72 native speakers that took part in the study.

As a conservative measure, the acceptability scores of only those shot types that each participant had watched for a total of at least 40 seconds are included in the analysis. To illustrate the differences in shot type mixes, Figure 36 presents the percentage at which a

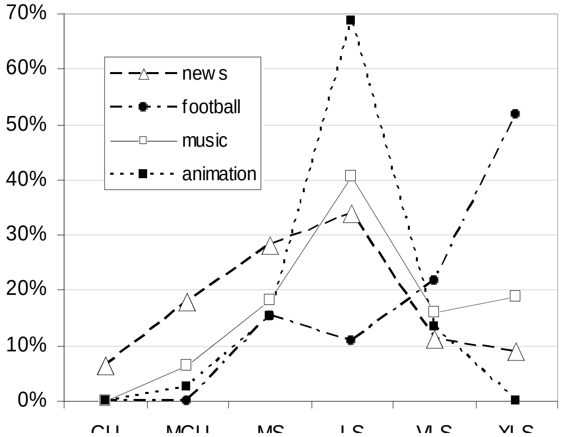


Figure 36: Distribution of shot type usage in experimental clips by content types

given shot type was used in the different content types. For example, roughly 50% of the football content was presented in extreme long shots, which were not used at all in the animation clips. *Shot type* was a significant predictor of acceptability [χ2(1)=148.4, p<0.001]. Averaged across all content types, sizes and encoding bitrates the *close up* and the *very long shot* were the most acceptable shot types. The *extreme long shot* (XLS) received the lowest ratings. All shot types became more acceptable with increased sizes (see Figure 37, left). The *extreme long shot* was by far the least acceptable shot at all sizes when averaged across the content types.

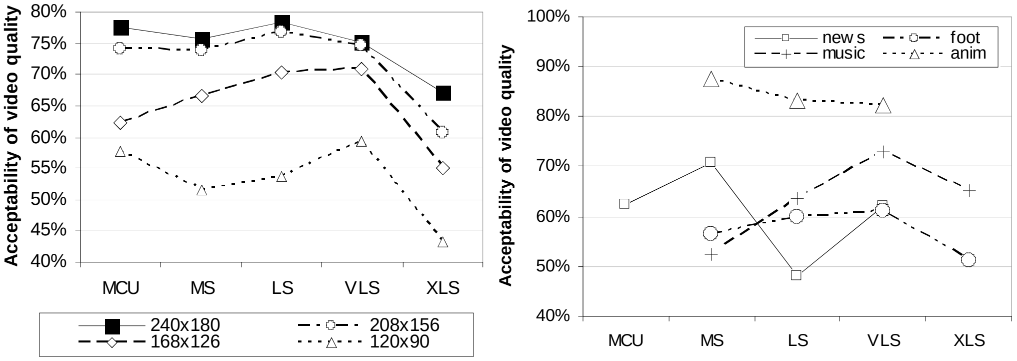


Figure 37: Acceptability of shot types by picture sizes (left) and by content types (right)

Furthermore, the regression revealed an interaction of *Shot Type* and *Content Type* [χ^2(1)=1337.1, p<0.001]. Figure 37 (right) illustrates the acceptability scores of the different shot types by content type averaged across the four different sizes and all encoding bitrates. The following subsection will address each content type in turn. For each size the acceptability scores of the shot type are averaged across all encoding bitrates. The figures below present these values with standard error bars based on the participants' acceptability averages in these conditions.

9.2.1 News

News content is made up of a mixture of different material and therefore had the biggest range of shot types in my experiment as can be seen in Figure 36. Typically the anchorman announced a topic that was then covered in more detail by means of field reports, graphs, illustrations or interviews. The field reports used a wide variety of shot types to depict the topic and to situate the audience. The video quality of the field reports was usually worse than the footage shot in the studio.

The shot type that yielded the highest acceptability of video quality for News across all sizes was the MS. One must keep in mind that this shot is typically used when presenting the anchor man in a static posture. The LS was the least acceptable shot type across all sizes. The acceptability of the video quality of the shot types at the two highest sizes did not differ significantly; Mann-Whitney [Z=-1.7, n.s.]. The acceptability of the different shot types is summarised in Figure 38 (left). The PSNR values of the different shot types presented in Figure 38 (right) looked very similar to the acceptability scores. The values for the MCU and VLS were about the same and the MS was slightly above and the LS slightly below in value. This provided evidence that for the news content the differences in acceptability between the shot types of a given size were merely due to differences in visual quality.

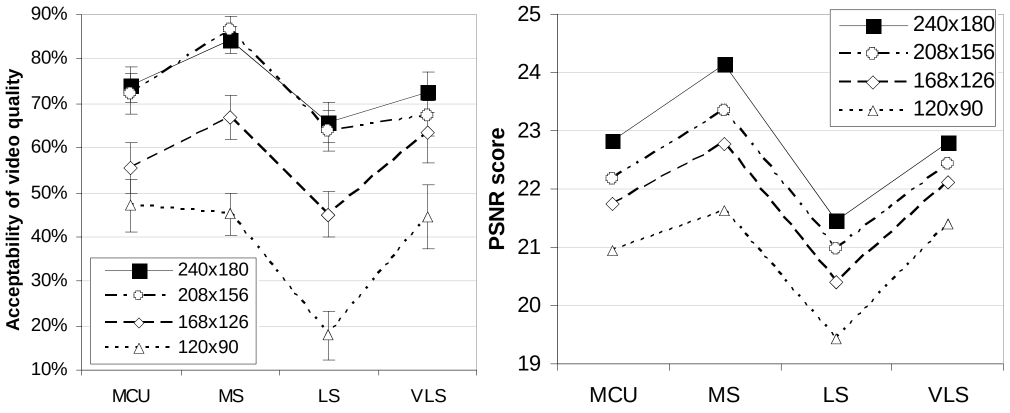


Figure 38: Acceptability (left) of shot types of News and corresponding PSNR scores (right)

9.2.2 Football

Almost all of the scenes in the football footage depicted players in motion or camera pans of the pitch. Shot types closer than a medium shot are not common in football coverage. It is hard to zoom in on and follow players because they often move in unpredictable ways. The extreme long shot provides the viewer with an overview of what is going on in the playing-field. It is very popular and even in the highlights material used in the study this shot was used approximately 50% of the time.

Non-parametric tests showed that there was no significant difference in acceptability of the extreme long shot at the largest size when compared to the other shot types [$\chi^2(3)=2.34$, *n.s.*]. However, at all sizes lower than 240x180 (VR>6.8) the results confirm the qualitative feedback about the extreme long shots in study 1. Here the extreme long shot was the least acceptable shot type. Surprisingly, the acceptability of the medium shot depicting the greatest amount of detail in the football material declined more than the long and the very long shot at smaller sizes (see Figure 39 left). However, this was only a trend and not a significant difference. In the computed PSNR values depicted in Figure 39 (right) one can find no evidence that the lower acceptability of XLS or MS might be induced by lower visual quality as was argued for the news content earlier. Both the MS and the XLS yielded considerably higher PSNR values in comparison to the LS and VLS.

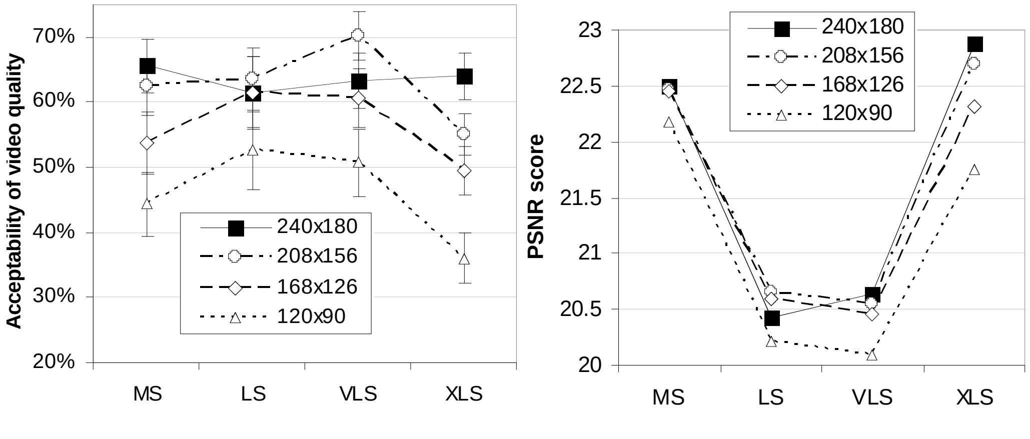


Figure 39: Acceptability (left) of shot types of Football and corresponding PSNR scores (right)

9.2.3 Music

The visuals of the music clips were dynamic with many camera pans. Across all sizes the medium shot was the least acceptable and the very long shot the most acceptable in the music clips. The acceptability of the less detailed shots (LS and VLS) increased with a corresponding decrease in the level of detail. The acceptability of the extreme long shot changed dramatically with different image sizes. At the smallest size its acceptability was only slightly above but not significantly different from the least acceptable medium shot. At the largest size, however, it was only slightly below and not significantly different from the most acceptable shot type – the very long shot (see Figure 40, left).

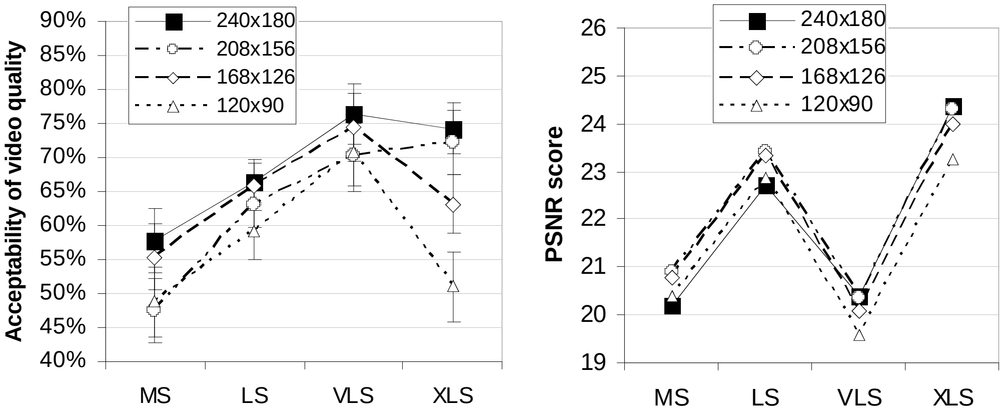


Figure 40: Acceptability (left) of shot types of Music and corresponding PSNR scores (right)

Apart from the XLS image size seemed to have little effect on the acceptability of the more detailed shots. One could neither explain the VLS's high acceptability across all sizes nor the XLS's reduction in acceptability at smaller sizes with just differences in visual quality. Figure 40 (right) shows that the VLS had the lowest PSNR scores of all shot types and they were close to the PSNR scores of the MS. Despite the low PSNR scores the acceptability of the VLS was the highest of all shot types for the music clips. The PSNR values provide no indication of the degradation of the XLS at smaller sizes that was evident from the acceptability scores.

9.2.4 Animation

The animation content relied mainly on three shot types: VLS, LS and MS. Shots with more detail than the medium shot might not be desirable because the imperfections of the claymation process, e.g. fingerprints, might become more visible. The animation content depicted fairly static scenes with few camera pans. Of all content types this was the easiest for the encoder to encode as can be derived from the PSNR scores, which are the highest of all the four content types (see Figure 41, right). In the fairly static animation content the medium shot presenting the most visual detail (MS) was the most acceptable. There were no significant differences between the long and very long shot in terms of acceptability.

The PSNR scores for the shot types of animation content depicted in Figure 41 (right) showed that the visual quality of the MS was the best and of the LS was the worst. The scores of the VLS lay between these two. The PSNR quality differences between the LS and the VLS were not reflected in the subjective acceptability values presented in Figure 41 (left) where LS and VLS were almost the same.

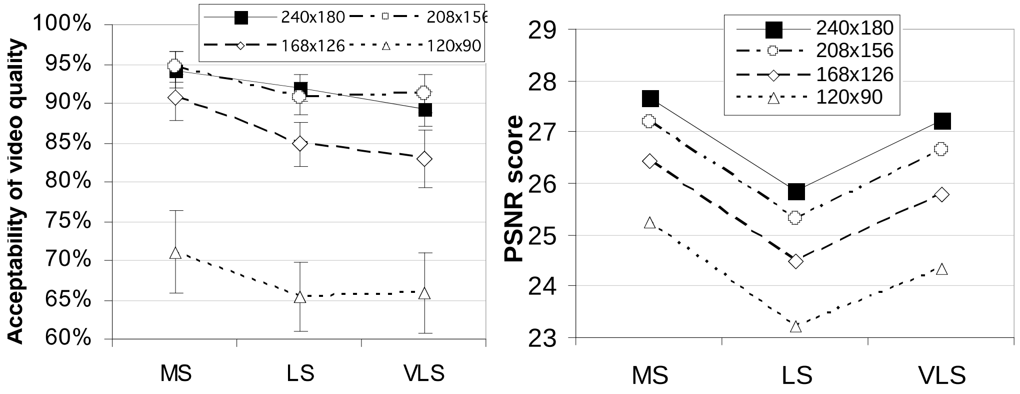


Figure 41: Acceptability (left) of shot types of Animation and corresponding PSNR scores (right)

9.3 Discussion

The acceptability of the extreme long shot declined most at sizes of 208x156 and lower in both music and football content. The acceptability of the very long shot, which shows a little more detail than the extreme long shot, was not degraded as much by these smaller sizes. This is encouraging news for cropping or zooming approaches (Dal Lago 2006), (Holmstrom 2003) that zoom in on part of the footage. Zooming brings the depicted content of an extreme long shot closer to what is seen in a very long shot, which had a much higher acceptability at all sizes lower than 240x180. More research is required to evaluate the potential benefits of cropping for sizes of 240x180 and higher, e.g. 320x240 that is supported by DVB-H (ETSI 2005).

The medium shot received the worst ratings of all shot types in the music clips. In the football clips only the extreme long shots received worse ratings. There are many possible reasons for this. Compared to the animation and news clips both of the former had many camera pans with moving background. For example, a football player is usually not static in this shot type. But camera pans were also used in other shot types both in football and music clips. One possible explanation is that in the medium shots the lack of detail due to the resolution and low encoding bitrates was most apparent. The unmet expectations of what should be visible in this kind of shot might also be responsible for low acceptability ratings. The importance of visual detail had also been noted in (McCarthy *et al.* 2004a) which found that visual detail was more important in football coverage than a smooth frame rate. The results at hand suggest that for sizes below 240x180 content producers should not favour the medium shot over other shot types when encoding bitrate is constrained and the subjects are in motion. Furthermore, this study found no support for relying on more detailed shot types such as the medium close-up to improve the viewing experience as reported in (Guardian 2005). The medium shot was not significantly less acceptable but at most sizes more acceptable than the medium close-up shot in the News content.

XLS were particularly affected by smaller sizes and the research community has been working on content adaptation that employs zooming into content for better depictions on small mobile screens. Guidelines for the amount of zoom were not existent and are the subject of studies 5a to 5c.

Chapter 10

Pick battles big enough to matter, small enough to win.
- K. Sozol

Study 5
Enhancing mobile TV through zooming

The results of study 4 in presented in Chapter 9 and other studies showed that

1) the perceived quality of TV material shown at small image sizes on mobile devices varies depending on the content e.g. (Song *et al.* 2004), and
2) not all shot types were equally affected by the depiction at smaller sizes at equal angular resolution.

Content providers like ESPN (Gwinn & Hughlett 2005), have identified XLS, which depict objects from a great distance, as the most problematic on small size screens. Since bespoke editing and content creation to minimize the use of XLS will be too expensive (Sylvers 2007) service providers need to repurpose existing TV content and adapt it for mobile TV through inexpensive means. The most promising approach to enhancing XLS on mobile devices is zooming, i.e., showing only part of the original footage and thereby offering larger sizes and higher level of detail of small objects. A number of technical solutions have already taken on the problem of automatically detecting scene boundaries (Lienhart 1999), identifying shot types (Voldhaug *et al.* 2005), choosing areas of interest within a frame (Agarwal *et al.* 2003) and suggesting dynamically zooming in and out of content (Sinha & Agarwal 2005). However, to date there has been no empirical research on what degrees of zoom for which target size result in the best experience for the user. The aim of the research presented in this chapter is to determine the size of objects and their level of detail users require in sports XLS, and how much of the surrounding contextual information can be traded off for detail to achieve the best overall experience.

The three studies presented in this chapter evaluated the most beneficial zooms for XLS in football content. Team sports, such as football, are good examples of resource demanding content with great audience appeal (*cf.* Sec. 5.2 on focus groups results). They investigated, which degree of zoom is preferred at which size. In addition to these subjective assessments, eye-tracking was employed as an objective way of monitoring user preference. The eye-tracking data allowed for an objective scene-by-scene comparison of the two different formats. Verbal feedback on the experience from 84 participants was collected through semi-structured interviews. The results are based on conservative estimates, which can be applied to other content types employing XLS.

Study 5a was designed to put the idea of zooming to the test for this type of content. It investigated the impact of a high zoom at natively depicted QVGA (320x240) resolution, the highest resolution currently targeted by mobile TV broadcasting formats (e.g. DVB-H), and QCIF (176x144), the smallest resolution. In study 5b I extended 5a by using two smaller zooms and three sizes at constant angular resolution to come to a better understanding of the limits of zooming. Finally, in study 5c I aimed at disambiguating the results in terms of size and resolution, which were confounded in study 5a and 5b, by letting people increase the size of zoomed and non-zoomed video clips up to their preferred size.

10.1 Study 5a

The method used in this study was to give the participants a choice of following a clip that was presented in two windows in parallel at different zooms (see Figure 43) by watching either the left or the right half of a screen. Since it only took a viewer a fraction of a second to change focus, this setup allowed for a very low cost switching between the presented video clips in terms of time and attention. This method interferes much less with the activity of watching TV content then any other of the methods discussed in (Nemethova *et al.* 2005). This study tested the idea of zooming with arelative high zoom (1.6) at the two extremes of the mobile resolution spectrum - for mobile TV services at the time the highest supported resolution was QVGA (320x240) and the lowest e.g. in DVB-H standards was QCIF (176x144). Besides researching the subjective preferences for the zoomed material, Eye-tracking technology was used to gather objective data on the participants' viewing pattern and individual interviews to obtain qualitative data.

10.1.1 Method

To find out whether participants preferred the zoomed material, I reviewed techniques used in video quality assessment. To assess gains or differences in quality between two versions of a video clip one can

1. present them sequentially one after the other,
2. display them side by side on one or more screens or
3. present one clip at a time but let the participant toggle back and forth between them by means of an input device.

These approaches have various advantages and disadvantages, which are discussed elsewhere (Stanger 2006). For this study I chose the side-by-side approach on one screen, which allowed for subtle differences in video quality to be detectable and a very low involvement of the participants in terms of head movement and required feedback. All of the material was presented as a choice between video clips.

10.1.2 Material

Football footage was recorded through freeview DVB-T at 758x576 and prepared for the clips. The first step de-interlaced the material, cropped off surrounding black bars and adapted the aspect ratio of the content to 4:3. I used Virtualdub's Lanczos3 filter to resize the material to 640x480. During this process the MSU LogoRemover filter removed text, i.e. the score of the game. This measure was motivated by the results from study 2, which had shown that text legibility had a major influence on the acceptability of overall video. Furthermore, the moving zoom window moved in and out of regions where the original score was present and this created a distraction when following the footage. These steps resulted in uncompressed source footage without text at a size of 640x480. Based on this a zoomed and a non-zoomed version of the material was produced. To create the non-zoomed version the base footage was resized to the two final sizes using a custom built C++ application written by Marco Papaleo. For the zoomed version the footage was screened frame by frame by Marco Papaleo and a 400x300 area of the 640x480 footage displaying the most important part of the extreme long shot was selected. Furthermore, the moving zoom window avoided to introduce unnecessary pans, which degrade the viewing experience (Holmstrom 2003). The area surrounding the zoom window was cropped off and the video was resized to the final size using the aforementioned C++ application. All other shot types remained unchanged and were identical for both resulting clips. To these zoomed and non-zoomed clips without text a current score text using Virtualdub's logo filter was added. For the QCIF size clips these scores used abbreviations of the club names (see Figure 42). The score had the same pixel size for the zoomed and the non-zoomed clips at the same size. Virtualdub logo filter was used to superimpose the current score of the game and compressed the resulting clips at 384kbps with Microsoft's MPEG4 V3 for the video and the audio at 16 bit PCM. In order to better understand the difference between the two resulting clips Figure 42 includes example screen shots depicting the same scene for the zoomed and non-zoomed clip. Within the used clips XLS were used 59% of the time. LS and MS made up the remaining time.

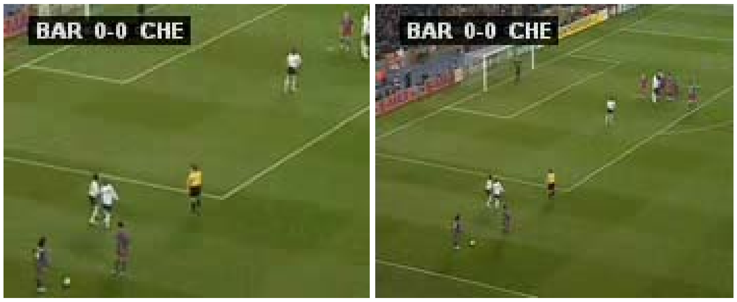


Figure 42: Zoomed (left) and non-zoomed material (right) with a zoom of 1.6 at 176x144

In order to present the two clips in synch a single file was generated that included both video clips and one of the (identical) audio tracks. A black clip in the middle spaced the two video clips 344 pixels apart for both sizes. In order to ensure that the clips were played at their nominal size on the screen when using media players full screen mode, black padding clips were created that were used on the left and right end of the screen. Avisynth's StackHorizontal function was used to create the final clip that had a total horizontal size of 1024 pixels. The dimensions of the video clips and their resulting angular sizes are summarised in Table 16.

10.1.3 Participants

Thirty-three paid participants (11f, 22m) took part in this study. Their average age was 29 years. The visual acuity was 100% for 30 of the participants, 95% (1), 85% (1), 80% (1). All of them were interested in football.

Table 16: Resolution and dimensions of content on the screen

Resolution	*Width*	*Height*	*VR*	*angular size*	*angular resolution*†
176x144	52mm	43mm	14 *H*	4.1º	35 ppd
320x240	94mm	71mm	8.5 *H*	6.8º	35 ppd

† To which degree the encoding bitrate achieved this nominal angular resolution is unknown.

10.1.4 Procedure

To control for possible effects due to imperfect visual acuity, the experimenter asked the participants to take a two-eyed Snellen test (Bennett 1965). After calibrating the eye-tracker, the participants watched two clip pairs, one of each size. The instructions stated that the participants could watch either one of the clips on the screen and could switch back and forth between the clips as many times as they wanted to. Both clips lasted for at least three and a half minutes. The monitor had a resolution of 1024x768 (with 86ppi). The participants followed the clips at a viewing distance of approximately 60cm, which is a more than the typical viewing distance of mobile TV consumption observed so far. But this way the setup matched the angular resolution of the presented material of studies 1 and 2. People with 100% visual acuity could discriminate all pixels at this distance (see Table 16 for the dimensions). The experimental design was counterbalanced in terms of size, left and right presentation of the zoomed footage. The chronological order of the content was judged more important for the ecological validity of the study than eradicating possible ordering effects. After each clip, the participants called out which clip they had preferred. For the first clip, there was an intermission of 15 seconds for this purpose. After the clips had played the experimenter asked the participants in the form of a semi-structured interview about their experience, and why exactly they had chosen one clip over the other (see Section A 4 for the questions).

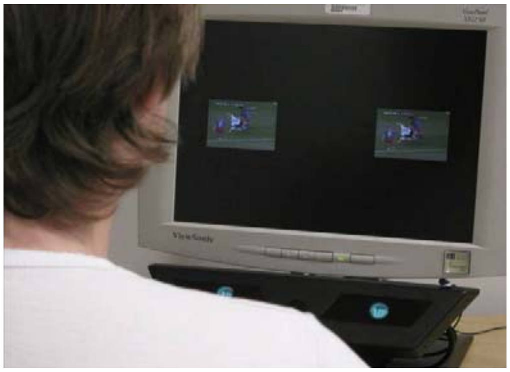


Figure 43: The participant's gaze while watching was captured through eye-tracking.

10.1.5 Results

I carried out non-parametrical Mann-Whitney tests on the participants' *visual acuity* and *gender* with respect to *preference. Visual acuity* denoted whether or not the participant's visual acuity was at least 100%. I found no significant differences for the *preference* of zoomed content due to *gender* or *visual acuity*. I averaged the binary *preference* data for zoomed over non-zoomed content for the two *sizes* across the participants. At QCIF *size* 61% of participants preferred the 1.6x zoomed content over the original content. For native QVGA depiction, only 24% of participants showed a *preference* for the zoomed material. A non-parametrical Wilcoxon test confirmed that the difference in *preference* between *sizes* was significant [z=-3.317, N=11, p<.001].

To analyze the eye-tracking data, a frame-by-frame analysis of the content assigned each frame its shot type. These shot type tags were then aligned to the matching eye-tracking data based on time stamps for each participant. All subsequent analysis is based solely on the XLS (59% of the total clip time), since all of the other footage was identical in the left and the right clips. Only the eye-tracking data from participants whose gaze had been captured during the majority of both the high and low size clips was used. The eye-tracking data showed that the participants followed the QCIF zoomed material at an average of 50% of the time during the experiment. This ratio dropped to 38% when the same content was presented at QVGA. The participants' *averaged time* for which they watched the zoomed content was not significantly different for the two *sizes* according to a paired T-test [t(23)=1.793; p=0.086]. Figure 44 summarizes both the subjective preference data and the relative amount of time spent watching the zoomed material at the two sizes.

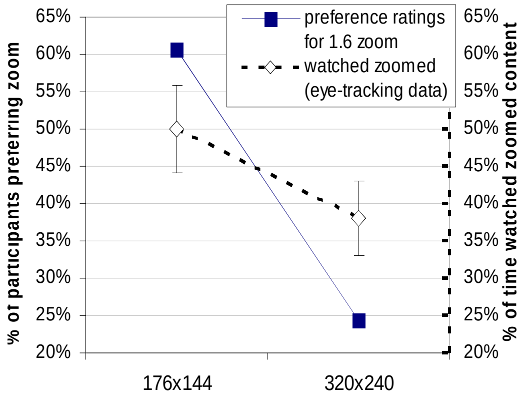


Figure 44: Preferences for zoom and percentage of time watched (with standard error bars)

The post-experimental interviews revealed that 55% of the participants based their choice on the perceived visual quality of the video, in particular for the QVGA clips. The participants' biggest complaint (*cf.* Figure 45) was about the visual quality of the zoomed material at QVGA, which they described as *'fuzzy'*, '*grainy'* or '*blurry'*. The non-zoomed material was described as '*crisp'* and '*clear'*. One participant's quote summarized this complaint: "[the zoomed in content]... *looked like a blurry podcast, if I'm close I want to see more detail". *A quarter of participants (25%), however, were not deterred by the reduced quality for the QVGA zoomed material, and preferred it for its larger depiction of the player and the ball, which they found easier to follow. For the smaller QCIF clips, few participants found the visual quality of the zoomed footage inferior to the non-zoomed. For QCIF, a majority of 61% preferred the zoomed clip. The most frequently given reason for watching the zoomed footage especially at QCIF was *'not being able to recognize the players'* or '*not able to see the ball'* in the non-zoomed clip. Previous research has found that recognizing players in football content is very important (McCarthy *et al.* 2004a). Twenty-four percent of participants made reference to the effort they had to put into following the non-zoomed clips at QCIF. '*Squinting'*, *'having to concentrate*' or *'looking hard'* were common complaints about the non-zoomed material. The participants that were opposed to zooming even for the QCIF were keen to be able to see as much as possible on the screen. Many participants made use of the zoom when they wanted to see a player in a tackle in more detail, or wanted to be able to see the players' feet and the ball properly. Some people mentioned that it would be nice to have both views accessible. In the interviews I asked the participants if they had looked back to the other clip after they had selected their preferred clip. Participants gave several reasons for switching away from their chosen clip. First, the situation was novel and many wanted to make sure that the quality of the clip they were not following had not changed, especially when close-up shots came up. This might have partly been instigated by the fact that the visual indeed changed since only the XLS had the zoom applied and the rest of the material looked identical. Many participants reported switching away from the main clip during close-up shots. Participants also switched to the non-zoomed view for more overview during long passes, corners, free kicks and crosses

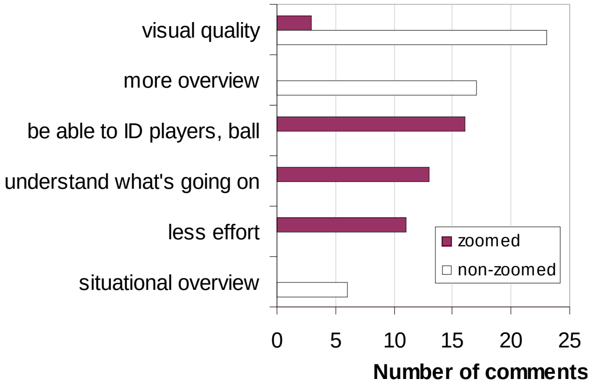


Figure 45: Reasons to watch preferred video (zoomed or non-zoomed)

whereas they switched to the zoomed view for tackles, dribbles or actions in the goal box in which they deemed detail of the players more important.

10.1.6 Discussion

Clearly the main reason participants preferred the non-zoomed content for the larger size clips was the lack of perceived quality at which the enlarged information was presented. It should be kept in mind that the footage was not up-sampled in size but that the presented footage was based on more pixels in the original albeit interlaced footage than what was presented on screen. Officials from the production company antena3TV confirmed that the resolution of the camera used in this game was 756x591. If I consider that the original TV footage's resolution was not the nominal 756x591 but had to be adjusted with a Kell factor of 0.6 the original footage had a resolution of 453x355. Consequently the 400x300 zoom (1.6) window of the 640x480 version of this clip only had an underlying resolution of 284x222 pixels. This means that the zoomed footage shown in the 320x240 format only had a resolution of 284x222. This results in an angular resolution of 32.8 ppd in comparison to the non-zoomed material, which with a resolution of 320x240 had an angular resolution of 35.4 ppd. For comparison Figure 46 includes a magnification of a frame of both the zoomed and non-zoomed QVGA footage.

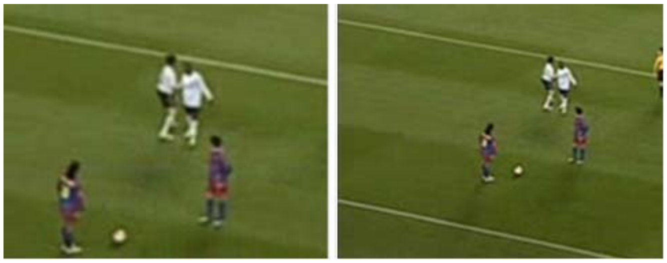


Figure 46: Magnification of the 320x240 XLS frame from Figure 42; zoomed (1.6) left, non-zoomed on the right

In comparison to the subjective preference, the eye-tracking data showed the same trend, but the differences between the two sizes were not significant. For the larger image size, the zoomed material was preferred significantly less then it was at the smaller size. The qualitative data told the clearest story as people complained about blurriness and lack of detail in the QVGA material. The quantitative data could not distinguish between the effects of this poor video quality and the reduced context for the zoomed-in – less of the pitch was visible - content at QVGA. In any case, high zooms (as 'low' as 1.6) of standard definition TV material were identified as potentially problematic, depending on the target resolution. Despite the significantly reduced context of the zoomed material, 61% of the participants preferred this over the original material at QCIF size. Considering participants' feedback on when they preferred the zoomed clips, it is evident that current zoom solutions e.g. (Sinha & Agarwal 2005) have to be far more sophisticated to match users' preferences on situational zooms. A high number of participants traded off viewing comfort for being able to see more of the available context in the QCIF size. It appears to be a good idea to provide services with zooming facilities that can be configured to users' preferences. Considering the many reasons participants had for switching away from their preferred clip, it is not surprising that the percentage of time watching the preferred clip was not as clear cut as the subjective preference data.

10.2 Study 5b

Due to the adverse effects of the 1.6 zoom on the participants' perception of video quality on the larger clips, a follow-up study tested two smaller zooms on three sizes.

10.2.1 Material

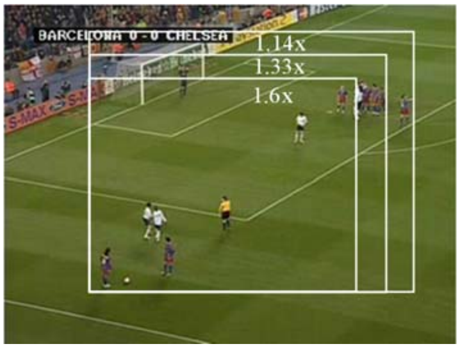


Figure 47: Example zoom areas in study 5b (1.14 and 1.33) and study 5a (1.6)

This study used exactly the same base material as study 5a, but used videos with two zoom levels. The zoom windows were 480x360 (1.33 zoom) and 560x420 (1.14 zoom) as illustrated in Figure 47. These were generated at three sizes: 176x144, 240x180 and 320x240. The dimensions of the 240x180 size clips on the screen were 71*mm* (height), 53*mm* (width) resulting in a viewing ratio of 11.3 *H* and an angular size of 5º at a viewing distance of 60cm. Figure 47 depicts an example frame with the zoomed areas of the different zoom factors from this stud (1.14 and 1.33) and study 5a (1.6). As depicted in Figure 47 the zoom areas were asymmetric with respect to the highest zoom and the centre of the original base material. The part of the picture that contained the most important information was identified through discussions between Marco Papaleo and me. The encoding bitrates of both audio and video were identical to study 5a.

10.2.2 Procedure

The procedure was identical to study 5a, except that the material was divided into three presentations, each of which was at least two and a half minutes long. Each presentation consisted of a zoomed and a non-zoomed video of identical size, which the participants watched in parallel on the screen (*cf.* Figure 43 on page 131). The zoomed clips featured the zoomed material for the parts in which XLS appeared whereas the non-zoomed clip showed the matching original material. The two clips differed only for the parts in which XLS appeared - all other shots were identical.

The participants were eye-tracked throughout the session. After each clip, the participants called out which clip (left or right) they had preferred. The factorial design of the experiment was counterbalanced for the order of the three different sizes of the clips and left-right occurrences of the zoomed clips. The independent variable *size* was tested within-subjects and the variable *zoom factor* between-subjects. The same interview as in study 5a at the end of the experiment, but it included additional questions to determine whether people had perceived any differences in quality.

10.2.3 Participants

The study included 51 paid participants (11 women, 40 men) with an average age of 29 years. Their visual acuity was 100% for 30 participants, 105% (5), 95% (6), 90% (4), 85% (1), 80% (4). All of them were interested in football.

Table 17: Resolution and dimensions of content on the screen

Resolution	*Width*	*Height*	*VR*	*angular size*	*angular resolution*[†]
176x144	52mm	43*mm*	14 *H*	4.1°	35 ppd
240x180	71mm	53*mm*	11.3 *H*	5°	35 ppd
320x240	94mm	71*mm*	8.5 *H*	6.8°	35 ppd

[†] To which degree the encoding bitrate achieved this nominal angular resolution is unknown.

10.2.4 Results

The dependent variable *preference* denoted whether or not participants preferred the zoomed material over the non-zoomed material. The binary preference replies from the participants for zoomed content for the three *sizes* and the two *zoom factors* were then averaged. As one might expect, *preference* for the zoomed content increased with decreasing size of the clips. For the smallest size more than 80% of the participants preferred the zoomed clips at their respective zooms. At the bigger *size* participants' *preference* for zoomed content decreased especially for the group with the 1.33 zoom. These results are summarized in Figure 48. The binary preference data was then analyzed through a binary logistic regression to test for main effects and interactions of the independent variables *zoom factor* and *size* on the dichotomous variable *preference*. The control variables *gender and visual acuity* were included. The latter denoted whether the participant had a visual acuity of at least 100%.

As in study 5a neither *gender* [$\chi^2(1)$=0.297; n.s] nor *visual acuity* [$\chi^2(1)$=0.969, n.s.] turned out to be significant predictors for the *preference* of zoomed material. The regression confirmed that *size* was a significant predictor for the participants' *preference* for zoomed content [$\chi^2(1)$=7.68, p<0.01]. Neither the *zoom factor* nor the interaction between the two independent variables turned out to be a significant predictor. In Figure 48 the results of the participants' preference are shown.

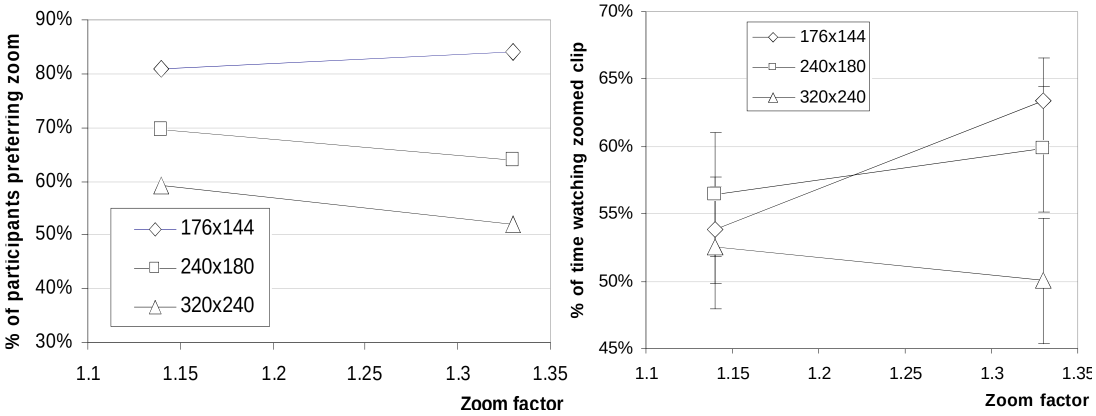


Figure 48: Preferences for zoomed content at different sizes and zooms (left), **the percentage of time zoomed content was watched** (right)

A trend of habituation was found: over time, participants' preference for the zoomed material increased. For the first two minute clip, the average preference for the zoomed content was only 57%, which rose to 71% for the third clip. Introduced into the regression analysis, however, this parameter turned out not to be a significant predictor for the participants preference for watching the zoomed footage

[$\chi^2(1)=2.42$;n.s.]. Considering this result and the fact that mobile TV interactions typically last 5-10 minutes it was decided to keep all of the existing data in the analysis for greater validity.

The analysis of the eye-tracking data showed similar results for the smallest and largest sizes. Their trends followed the preference data. At QCIF size the 1.33 zoom clips were followed 54% of the time and at the 1.14 zoom 64% of the time. At the QVGA size the zoomed content was followed less at the 1.33 zoom (50%) than at the 1.14 zoom (52%). The percentages of time people watched the zoomed material at the 240x180 size clips had a trend in the opposite direction of the preference data. At the 1.14 zoom participants watched the zoomed clips 56% whereas the group with the 1.33 zoom followed it 60% of the time. This difference, however, was not significant. A two factor mixed design ANOVA showed a significant effect for *size* on the dependent variable *time watching zoomed content* F(2,86)=3.261;p<.05. Neither *zoom* nor the interaction of *zoom* and *size* turned out to be significant.

The higher number of people preferring zoomed clips in comparison to study 5a was also mirrored by the qualitative feedback obtained in the interviews. The participants made 107 comments on the criteria on which they had based their choice. The summary of the most frequent reasons is presented in Figure 49. The visual quality was again a very important criterion. However, in this study with smaller zooms most participants deemed the zoomed material of better quality in general – many described it as '*clearer'*. This was not unanimous, however. Nine comments described the zoomed content as '*blurry'* and the non-zoomed material as '*clearer'*. The most important reason for not watching the zoomed material was that participants wanted to see more of the pitch in general or in specific situations like corners, passes and free kicks. The participants who preferred the zoomed material said it was more comfortable to follow and required less effort. People watched it to see the players and follow the ball better. They preferred to be closer to the action in general and specifically in tackle, dribble and goal box situations. In accordance with the preference data many participants that favoured overview over detail and viewing comfort said that at smaller and especially the smallest size they preferred the zoomed material as the non-zoomed material was too small and hard to watch.

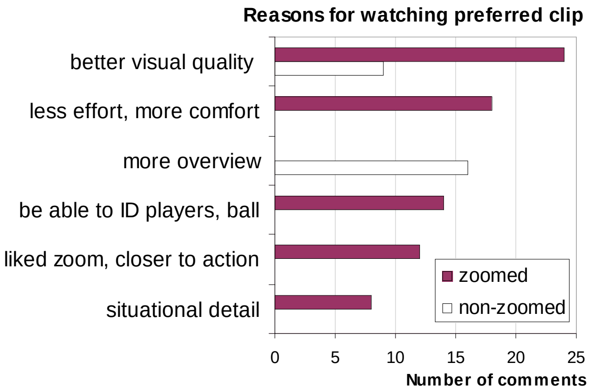


Figure 49: Reasons for watching preferred clip

10.3 Discussion

Across both studies the results showed that a majority of the participants preferred the zoomed content when presented at a size lower then 320x240. Even at the largest size a majority of participants preferred a small zoom (1.14) over the original. Since both studies were conducted in the same way and were based on the same footage a binary logistic regression was performed on the combined preference data from study 5a and 5b. As in both individual analyses size was a significant predictor of preference [χ2(1)=12.75; p<0.001]. The regression also revealed zoom factor as a significant predictor for the preference for zoomed content [χ2(1)=16.002; p<0.001]. The interaction between the two independent

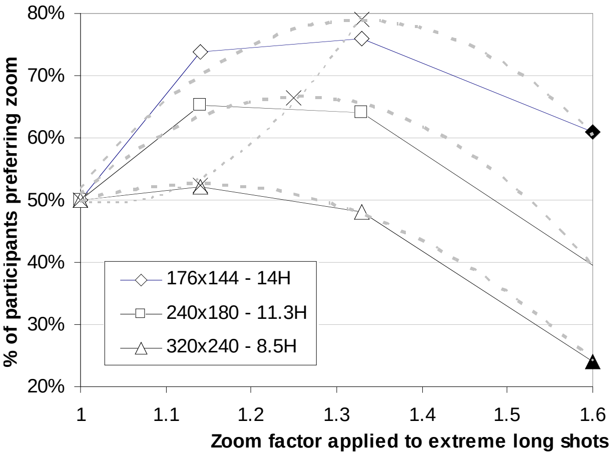


Figure 50: Combined data – study 5a (black), study 5b (white) and assumed origin (zoom factor 1x).

variables was not a significant predictor for preferring the zoomed content. I combined the preference results from study 5a and 5b in Figure 50. The optimal zooms are marked with an X. Additionally, the graphs include an assumed 50% chance preference if the zoomed content was identical to the non-zoomed material at a zoom factor of 1. Last but not least the graph includes an interpolated value for the 240x180 size for the zoom factor 1.6. Based on these graphs second order polynomial trend lines (dotted) were computed for each size, all of which had an R^2 of at least 0.94. From these graphs one can derive the preferred zooms for XLS for the three sizes: 320x240 (1.17), 240x180 (1.26) and 176x144 (1.32). The zoom factors for these values can be approximated (R^2=0.99) with the following linear function that uses the height in pixels as input: f(x)=-0.0015x + 1.53. This function should return the optimal zoom factor for all sizes that are within the studied range of study 5a and 5b. It should be kept in mind that the zooming window in these two experiments was based on human decision making - a best possible case - and that automated approaches might result in sub-optimally cropping off material that includes useful context information.

Using eye-tracking results as an indicator for preference was not as clear cut as the data about participants' preferences. The participants reported that they switched back and forth many times, for comparison reasons, out of curiosity, novelty of the setup, reassurance that they made the right decision, induced by camera pans or action going from left to the right on the screen or vice versa. Considering the participants' various mentions of the recognition of players as their main criteria I sampled typical sizes of players on the screen in extreme long shots who were close to the action. The average in the original VGA material was 42 pixels. With the above optimal zooms this results in players sizes from 17 pixels for QCIF to 25 pixels in height for the QVGA resolution when presented at 35ppd.

10.4 Study 5c

This study was performed as a part of study 6 but it is part of the evaluation of size requirements in zooming approaches. It focuses on the influence of *actor size* on the acceptability and the preferred viewing conditions of sports XLS.

10.4.1 Material

I used 10 second parts of two video scenes of which Marco Papaleo had produced a zoomed and non-zoomed version in the context of study 5a. The two scenes depicted exclusively XLS of football from two different distances. Scaled down to 176x144, the player size averaged 12 and 15 pixels in the two non-zoomed clips. In the two zoomed versions a zoom factor of 1.6 resulted in average player height of 18 and 24 pixels when the clip was scaled down to 176x144.

This provided me with four different sizes of the actors in the footage 12, 15, 18 and 24 pixels in height. This would allow me to find out whether participants' preferences in terms of preferred size are due to the absolute size of the clips or depicted objects within the video clips. Originally all of these clips had been encoded at a resolution of 176x144 at 350kbps WMV V9 at 12.5 fps and WMA V8 at 32kbps. I encoded each of these four clips at six different dimensions 480x360, 400x300, 320x240, 240x180, 168x126 and 120x90 at all of them at the same higher encoding bitrate (1Mbit) in order to ensure that the resulting clips had the same visual quality. They would appear bigger on the screen but with the same resolution. The original pixels would now be stretched over more pixels on the display (*cf.* on page 32).

10.4.2 Apparatus, participants and procedure

The apparatus and participants were identical with study 6. Please refer to Sec. 11.2 and Sec. 11.3 respectively on page 142 for the details. After having watched 16 clips the participants saw four XLS clips in a randomized order. The presentation assured that the zoomed and non-zoomed version of the two base clips were not played back to back. In the debrief interview I asked participants to indicate whether or not they considered themselves football fans. Nineteen of the 36 participants fell into that category.

10.4.3 Results

I averaged the acceptability scores of all participants of the six picture heights and of the four clips to obtain the acceptability curves presented in Figure 51. XLS clips depicting actors that were larger in size – either through zoom or the fact that the original scene was closer to the players - were generally more acceptable at all sizes smaller than 37.5*mm* (8.5*H*, 6.7°). Once the viewing ratio reached around 8H the benefits of the zooms diminished – the four clips' acceptability scores were at similar levels. At viewing ratios larger than 14H even the clip with the largest depictions of actors achieved only an acceptability of 52% (60% for fans). The acceptability of all four clips reached its maximum at the two largest sizes (*cf.* Figure 51). This means that the measures *favourite size* and *minimal angular resolution* are subject to possible ceiling effects because no larger sizes could be selected by the participants. Since Kolmogorov-Smirnov tests showed that *favourite size*, *minimal* size and *minimal angular resolution* did not follow normal distributions; I performed non-parametric Friedman tests on these dependent variables. Smaller depictions of players made participants prefer larger sizes (χ^2=21.9, *df*=3; p<0.01). The mean *favourite size* increased from 38.5*mm* (18ppd) for the 24px to 42.5*mm* (17ppd) for 12px player heights. The matching favourite angular sizes for actors were between 1.3º and 0.7º. Analogously, for smaller player depictions people required larger *minimal sizes*

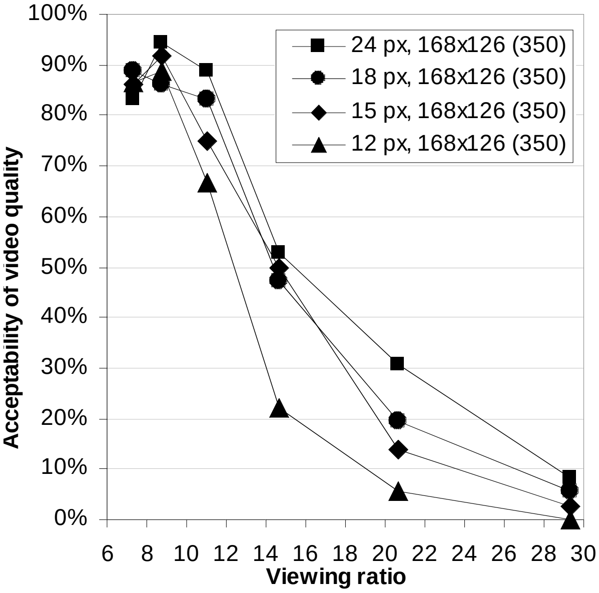


Figure 51: The acceptability of sports XLS clips by viewing ratio for different *actor sizes*

(χ^2=31.1, *df*=3; p<0.01). The angular size of the depicted actors for this lower bound ranged from 0.5º to 0.8º. There was no significant effect of *actor size* on the *minimal angular resolution* (χ^2=6.1, *df*=3; p=0.11). I received a number of qualitative comments from the participants who remarked that the quality of these football clips were higher than the ones they had seen in the previous 16 clips. The zoomed clips were encoded at 350kbps and the clips that were shown before (*cf.* Chapter 8) only at 192kbps. Football fans found the footage significantly more acceptable than non-fans at all sizes (t(36)=-3.1;p<0.01).

10.4.4 Discussion

Viewing ratios between 8 and a maximum of 11 should result in the best experience of sports QCIF content that includes XLS. In terms of acceptability the benefit of zooming diminished once the VR of the whole picture reached 8. At this viewing ratio large (1.3º) player depictions resulted in a visual quality better than acceptable as it reached the ceiling of the used scale. For VRs of 14 and larger there was still a large benefit for zooming but the overall acceptability (52%) was low. For a fair comparison with the results of study 5c with study 5a and 5b I considered only the ratings of football fans (*cf.* Figure 52) and inspected ratings for the VRs of 8 and 11 (see the grey box in Figure 52) in more detail in Figure 53. For an acceptable experience for all fans an angular player size between 0.7 º and 0.8º was required and the data collated in Figure 53 suggested that XLS acceptability could be thought of as a logistic function of player size for fans given a sufficiently high encoding bitrate. This matched with the favourite sizes in study 5c which for the smallest players (12px in Figure 51) at around 8H resulted in an angular size of 0.7º. Furthermore it aligns with the combined results of studies 5a and 5b in which the preferred size of actors in XLS was between 0.5º (176x144, 14H) and 0.7º (320x240, 8.5H). It seems that at a sufficiently small viewing ratio of 8.5 the participants would just allow for a very small zoom to attain 0.7º player size and at the same time to sacrifice the least of the pitch for the overview. Fans had to trade off the increased actor sizes for a reduction in visual context while the angular resolution of the picture was constant (35ppd). In study 5c increasing the size of players did not reduce context - only the angular resolution of the depicted video.

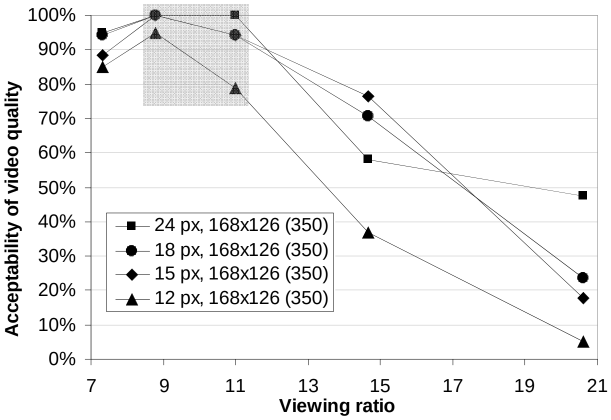


Figure 52: Fans' acceptability of XLS dependent on VR by player size in pixel

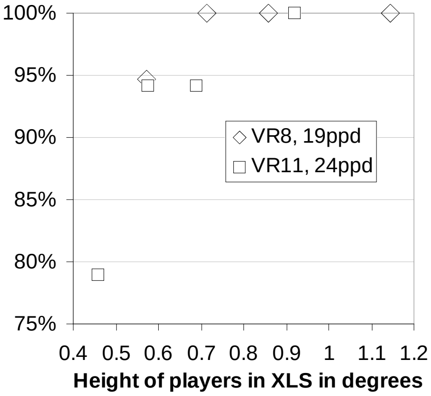


Figure 53: Fans' acceptability of XLS dependent on player size

10.5 Conclusion

Study 5c showed that player size is secondary to the overall picture size of QCIF content. Only once the picture was big enough (11H) did football content appeal to a large majority of participants. At this size (11H) the value of zooming was large – content with players of 0.9º angular size was acceptable to all football fans. Based on the data from 84 participants the studies revealed the preferred zooms for XLS of football content of standard digital TV resolution on the current mobile TV size range. A conservative acceptability threshold for the angular size of players was 0.8º. XLS presented at these optimal zooms appeared to have better visual quality to a majority of participants, required less effort to watch and made recognizing players and the ball easier.

The studies showed that zooms larger than 1.3 can have adverse effects on people's experience of SDTV footage at QVGA resolution. Objections to high zooms could be explained by the importance of the strategic information contained in the pitch when the VR is small enough (around 8) or by the reduced angular resolution. Assuming a combined Kell and interlace factor of 0.6 – a more conservative estimate than 0.7 taken in the rest of this document due to the large amount of motion - the actual resolution of the zoomed content was only 32.7ppd in comparison to 35.4ppd for the non-zoomed material. This restriction will apply to mobile content in general when content is captured with interlaced SDTV cameras.

The main limitation to this study is that my participants were quite young on average and for older people other zoom values might provide a better experience. Content that might be not as sensitive to loss of context as football is might have slightly higher optimal zooms. A more thorough discussion in conjunction with the results from study 6 is presented in Sec. 12.7. In the next chapter I will explore in detail how in general the trade-off between size and angular resolution depends on content types, shot types and content resolution.

Chapter 11

There are no solutions … there are only trade-offs
T. Sowell

Study 6

On the trade-off between resolution and size

In the results of study 1 presented in Chapter 6 size and resolution were confounded; it was not clear if the reduced acceptability of lower resolution clips was due to the size of the video being too small or the reduced amount of detail. In order to disambiguate the findings I conducted this follow-up study, which included shot types as another variable. In this study participants could decide to watch footage at six different sizes of video that were all based on the same base resolution clip. The clips were blown up to various extents on a high resolution device (200ppi). Participants chose their favourite size and also provided acceptability ratings on all six different sizes along with explanations why this was the case. The study evaluated participants' preferences for two different resolutions in a between-subject design and used the same content mix as study 1.

11.1 Material

In this study I re-used football, music and animation material already used in study 1 and the news material from study 2. The news clips had legible text at all sizes and the text was of the same quality as the video material, referred to as degrading text in Chapter 7. This material did not provide consecutive shots of all types lasting for more than ten seconds. So in order to control for effects due to shot types I used a consecutive shot of 8-10 seconds. Due to differences in content, the most detailed shot types were not completely identical. For example, the football shot with the most detail was closer to a mid-shot than a medium close-up, and the most detailed shot in animation was closer to a close-up. In general, all less detailed shot types were comparable but the XLS of football and news depicted people far away whereas the XLS of music (moving camera) and animation (static shot) depicted a landscape.

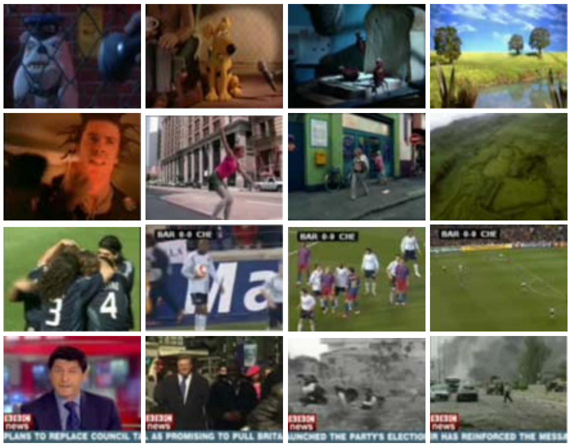

BAR 0-0 CHE
BAR 0-0 CHE
BAR 0-0 CHE
BBC news
PLANS TO REPLACE COUNCIL T
AS PROMISING TO PULL BRITA
UNCHED THE PARTY'S ELECTION
R HAS REINFORCED THE MESSA

Figure 54: Shot types of animation, music football and news f.l.t.r. MCU/MS, LS, VLS, XLS

The clips were originally encoded at 192kbps WMV V9 at 12.5 fps at two respective resolutions 120x90 and 168x126 with Audio V8 at 32kbps. The two sets of clips were then encoded at the six different dimensions 480x360, 400x300, 320x240, 240x180, 168x126 and 120x90 at a higher encoding bitrate in order to ensure that the resulting clips had the same visual quality. They appeared bigger on the screen but their resolution was the same. The original pixels would now be stretched over more pixels on the display. The text contained in the news clips was legible at all above sizes.

However, I wanted to control for possible effects due to shot types. In order to achieve this I had to use short clips as it was hard to find footage that used the same shot type consecutively for more than ten seconds. The dimensions and values for angular size, angular resolution and viewing ratio are summarized in Table 18.

11.2 Apparatus

The clips were presented on an iPAQ hx4700 with a 200ppi VGA resolution transflective TFT display with 64k colours. At a viewing distance of 32cm, this resulted in a nominal angular resolution of 45 ppd. As in studies 1 and 2 the iPAQ was equipped with a set of Sony MDR-Q66LW headphones to deliver the audio. Each set of the six different size clips was arranged in a play list and played through the application *The Core Pocket Media Player* (TCPMP version 0.71). I checked that all clips played at their nominal frame rate using the benchmarking tool included in the TCPMP. Benchmarking videos encoded at 640x480 pixels showed that videos did not play at their nominal frame rate of 12.5 fps. The highest resolution that would play at the nominal frame rate was 480x360 which was then chosen as the maximum for this study. The dimensions and values for angular size, angular resolution, and viewing ratio are summarized in Table 18.

11.3 Participants

A total of 35 paid participants (18f, 17m) with an average age of 25 took part in this study. Thirty participant had a visual acuity of 100% or better, 95% (1), 85% (1), 80% (1) according to a Snellen test. Two male participants were colour-blind according to an Ishihara test.

Table 18: Experimental setup values based on 32cm viewing distance

Video area	*Pixels used to display*		*VR* *	*Ang*	*AR ppd* *†	*AR ppd* *†
Width x height	*the video*			*Size**	*120x90*	*168x126*
60*mm* x 45mm	(480 x 360)	172,800	7.1	8º	11	16
50*mm* x 37.5mm	(400 x 300)	120,000	8.5	6.7º	13	19
40*mm* x 30mm	(320 x 240)	76,800	10.7	5.4º	17	23
30*mm* x 22.5m	(240 x 180)	43,200	14.2	4º	22	31
21*mm* x 16mm	(168 x 126)	21,268	20	2.9º	31	44
15*mm* x 11.25mm	(120 x 90)	10,800	28.4	2º	45	45‡

† To what the encoding bitrate achieved this nominal angular resolution is unknown.

‡ The resolution of this footage was only 120x90 limited by the resolution of the display.

* All the values are based on 32cm viewing distance the average observed in this experiment. *AR* and *VR* are rounded values.

11.4 Procedure

Prior to the experiment I tested participants for visual acuity with a Snellen chart, and for colour-blindness with an Ishihara test (see Sec. A 8). The participants watched the clips while seated on a couch in a lab with ambient light of 345 lux. The instructions stated that the participants could assume any position that they preferred and that they deemed appropriate for following mobile TV.

The participants watched 16 clips in a randomized order. But in each four clips played (1-4, 5-8, 9-12, 13-16), each content type would appear at least once and each shot type appeared at least once. The presentation ensured that each content type and shot type combination was used at least once as the first clip. Each clip started playing the smallest size and the participants were asked to find out their preferred size for this clip. They could use buttons to increase or decrease the size. On each button press, the video started from the beginning again.

The participants were told to find their *favourite size* in terms of the best visual experience for them and point out which sizes they deemed *acceptable* and *unacceptable* as a viewing experience for a mobile TV service. I encouraged and prompted the participants to provide reasons why they found certain sizes unacceptable, which resulted in them *complaining aloud* for each clip's presentation. Finding one's preferred size is akin to the *method of adjustment,* which was successfully adopted in previous video quality research by, e.g. Richardson *et al.* (2004). The method of acceptability was introduced in Sec. 3.5.1.

The participants and their interactions with the clips were audio and video recorded. Viewing distance measures were also taken by means of a measuring stick that was occasionally held at the side of the participants, which did not seem to interfere with the participants' task.

11.5 Results

For each video clip I obtained three measures - the *favourite size* at which participants preferred to watch, the *minimal size* and the *minimal angular resolution* (derived from the largest acceptable size) at which watching was still acceptable. I ran three mixed factor ANOVAs on *favourite size, minimal acceptable size* and *minimal angular resolution* as the dependent variables each with *content type* and *shot type* as *within-* and *resolution* as a *between-subjects* factor, which will be addressed separately below. Each ANOVA was based on 560 individual ratings. Angular sizes are reported in degrees, viewing ratios in terms of picture height (*H)* and angular resolutions in ppd.

11.5.1 Viewing distance

The obtained viewing distances were averaged for each participant. Based on these I computed descriptive statistics. The median and the average viewing distance were both 32cm with a standard deviation of 6.8cm. Although the average viewing distance in the 168x126 resolution group was slightly higher (32.7cm; σ =6cm) than in the 120x90 group (31.8cm, σ =7.6cm) a t-test showed that this difference was not significant: t(33)=-0.372, n.s. Only one participant systematically varied the viewing distance with the six different size videos – pulling it closer for the smaller images. All other participants generally assumed the same posture when flicking through the different sizes. When they were unsure about the acceptability of a small size clip they occasionally pulled it closer for inspection but then usually changed back into their preferred position.

11.5.2 Acceptability of the visual experience

I averaged the acceptability scores of all participants for the six different sizes in both resolution groups (depicted in Figure 55, left). The acceptability varied tremendously with the size of the video. The averaged acceptability values for both resolutions increased greatly for the larger sizes in comparison to the smallest size (picture height 11.25mm). However, the acceptability then reached a local maximum – 80% at 30*mm* picture height for the 120x90 resolution and 90% at 37.5*mm* picture height for the 168x126 resolution – after which the acceptability dropped off. The second order polynomial trend lines of the averaged acceptability scores were:

120x90: $y = -0.0016x^2 + 0.0988x - 0.6948$; $(R^2 = 0.985)$,

168x126: $y = -0.0015x^2 + 0.1075x - 1.0051$; $(R^2 = 0.973)$.

They result in local maxima of acceptability at a picture height of 31*mm* for 120x90 (10.3*H*, 5.5°, 16ppd) and 35.5*mm* for 168x126 (9*H*, 6.3°, 20ppd). In Figure 55 (right) the acceptability ratings are plotted dependent on the angular resolution. For angular resolutions higher than 20ppd curves seem to differ only by a constant offset with larger picture sizes resulting in higher acceptability. For viewing ratios 14 and 20 we can see that for a constant size decreasing the angular resolution resulted in higher acceptability.

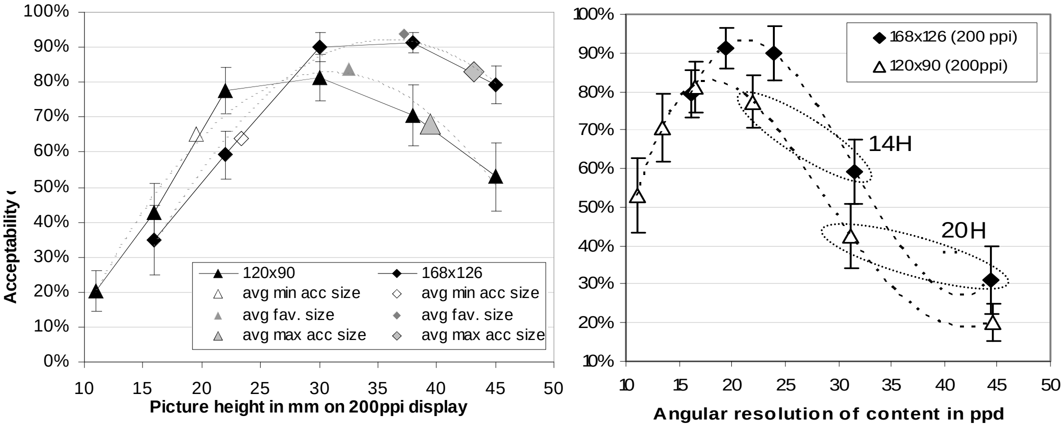


Figure 55: Acceptability of visual experience dependening on picture height (left) and angular resolution (right) by content resolution

11.5.3 Favourite size

The ANOVA showed that the *between-subjects* factor *resolution* had a significant effect on the participants' *favourite size* F(1,33)=5.47, p<0.05: the participants in the higher resolution group favoured larger sizes (37.2mm, 8.6*H*, 19ppd) than those who watched lower resolution material (32.6mm, 9.8*H*, 15ppd). This was a large effect size (Cohen's) *d*=0.75. The average favourite sizes of all participants of the two resolution groups were slightly larger than the computed maxima of the polynomial trend lines in Figure 55 based on the averaged acceptability results.

There was a significant main effect for *content type* *F*(3,99)=5.5, *p*<0.01. At closer inspection the Bonferroni adjusted pair comparisons showed that this effect was due to the news content in the low resolution group with an average *favourite size* of 30*mm* (10.5*H*, 17ppd) - significantly smaller than the 33*mm* of other content types (9.6*H*, 15ppd) in the low resolution group. In the higher resolution group the favourite size of news was not different from the mean of the other content types (8.6*H*, 19ppd).

No significant effect was found for *shot type*. The interaction between *content type* and *shot type* was significant (*F*(9,297)=3.35, *p*<.01) due to the football's XLS, which the participants preferred to watch at 39*mm* (8.2*H*, 18ppd) a significantly larger size compared to the XLS of animation and news at 35*mm* (9.1*H,* 20ppd).

11.5.4 Minimal size

Higher *resolution* content had to be presented at a larger size than lower *resolution* content in order to be equally acceptable (*cf.* Figure 55). For the high *resolution* video clips at 168x126 the *minimal acceptable size* was 23.4mm (VR=13.9) – significantly larger than the 19.6mm (VR=16.3) for the low *resolution* clips (*F*(1,32)=7.32 *p*<0.05). I found a significant main effect for *shot type* (*F*(1,32)=40.71, *p*<0.001). The average minimal acceptable size of the two more detailed shots was 19.5mm (VR=16.4) LS and 21mm (VR=15.2) for the MCU/MS significantly smaller than for XLS and VLS (both around 23mm, VR=13.9). An interaction effect between *shot type* and *resolution* (*F*(1,288)=10.78, *p*<0.001) (illustrated in Figure 56) showed that for the low *resolution* clips the differences between shot types as described in the main effect for *shot type* were smaller. The only difference that remained significant was the required minimal size for XLS (20.8mm) in comparison to the MCU/MS (18.2mm) for low resolution.

There was a significant effect for *content type* (*F*(1,32)=7.32 *p*<0.05) on *minimal acceptable size.* This was due to the football's XLS which required larger sizes (23mm, VR=13.9) than the XLS of the other content types (21mm, VR=15.2). Similarly, an interaction effect between *shot type* and *content type* was based on individual clip differences - the animation's VLS, a relatively dark shot, the news's LS with the presenter being occasionally occluded and the football's XLS. They all required larger sizes to be acceptable. The animation's static LS was acceptable at smaller sizes than the other LS shots. An interaction effect between *shot type* and *resolution* (*F*(1,288)=10.78, *p*<0.001) (illustrated in Figure 56) showed that for the high *resolution* clips the differences in *minimal acceptable size* due to *shot type*s were more pronounced. For the low *resolution* the only difference that remained significant was the required size for XLS (20.8mm) in comparison to the MCU/MS (18.2mm).

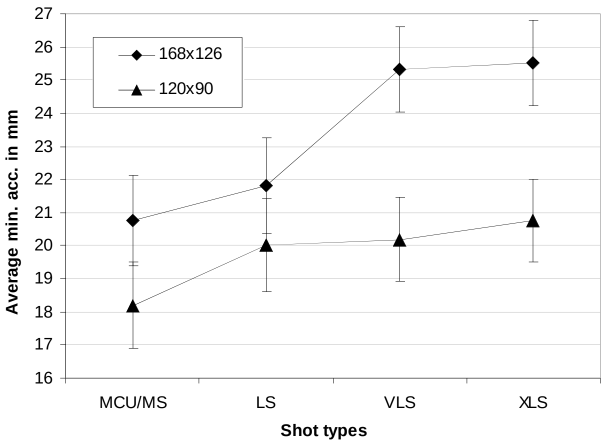


Figure 56: Interaction effect of *shot type* and *resolution* on *minimal acceptable size* (in mm)

11.5.5 Minimal angular resolution

Resolution was the only factor that had a significant effect on the acceptable *minimal angular resolution* (*F*(1,33)=7.05,p<.05). The average lower bound was larger for the 168x126 group (17ppd) than for the 120x90 group (13.5ppd). The corresponding average maximum picture heights were 43mm (σ =4mm) and 40mm (σ =7.5mm). I discuss the possibility of this being due to a ceiling effect in Sec 11.6.

11.5.6 Qualitative results

The qualitative results are based on 801 comments I collected. From the qualitative feedback I found that people deemed the smaller sizes unacceptable because they found them *"too small"*, "*couldn't figure out what's going on",* "*hard to identify people*" and *"hard to look at"*. The number of these complaints (depicted in Figure 57, left) dropped off once the size reached 30*mm* in height (11*H*, 5º). Some participants commented that although the definition seemed high - the image size was not big enough to appreciate it. With the larger image sizes, the experience was rated unacceptable because of the lack of definition or resolution. For both groups, complaints about definition started once the viewing ratio was 14 (equating to angular resolutions lower than 31ppd for the 168x126 and lower than 24ppd for the 120x90). Once the angular resolution fell below 20ppd (see Figure 57, right), the number of complaints increased dramatically. Lack of definition was a common complaint about text albeit to a lesser degree. With small image sizes (<22mm), participants complained about the effort required to read the text: with larger sizes and lower angular resolution (<17ppd), the quality of the text became too '*blurred*', '*pixelated*' or '*fuzzy*'. Other problems mentioned in connection with smaller images were dark scenes, insufficient contrast, and movement (either of the camera, or in the scene). For all angular resolutions lower than 24ppd (*cf.* Figure 57, right) the higher resolution group (which saw a larger picture than the lower resolution group but at the same angular resolution) made more complaints about insufficient definition than the low resolution group.

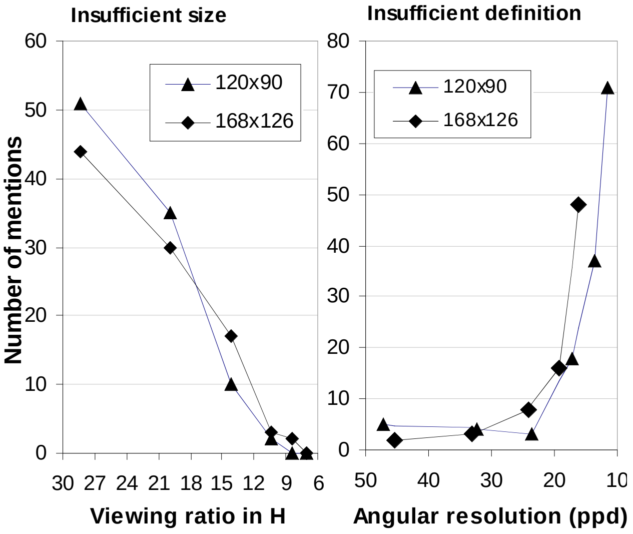


Figure 57: Participants' complaints about insufficient size (left) and insufficient definition (right)

11.6 Discussion

I will discuss the results of this study in relation to previous research in more detail in the following chapter.

11.6.1 Viewing distance

As in study 1 there was no significant change in viewing distance during the experiment. The average viewing distance of 32cm did not change over the course of the 16 clips watched during the experiment. The viewing distances observed in this study are similar to the 35cm found by Kato *et al.* (2005) and 28cm in study 1. I found no evidence that it depended on the size or the resolution of the displayed videos. The average viewing distance of 32cm observed in study 6 is similar the 35cm reported by Kato *et al.* obtained from people watching video on a 166ppi device.

The fact that the viewing distance observed in this trial was larger than in study 1 could be attributed to the following factors:

1. **Accuracy:** The measures reported in study 1 were obtained by estimating viewing distances, based on observational video recordings. In this study, the viewing distance was measured more directly.
2. **Addressability**: The higher resolution of the 200ppi display in comparison to the 115ppi display in study 1. In the previous study this resulted in an angular resolution of 21ppd for all viewing ratios, which ranged from 6.8 to 13.5.
3. **Method**: In the previous study the participants saw smaller picture sizes on average, had no control over the size of the clips, had to sit through the whole video and had to use a stylus to contribute their feedback. In this study they could quickly flick through with button presses and verbally discard sizes that they did not find acceptable.
4. **Furniture**: In this study the participants were told that they should assume a comfortable posture that they would assume if they were watching mobile TV. They were seated on a sofa rather than a chair with armrests, which might have affected their posture.

11.6.2 Minimal size

Shot types depicting objects from close up could be watched at smaller image sizes. Similar to the results on favourite size (see below), higher resolution required larger sizes to be acceptable. More research is required to explore the full extent of the interaction between resolution and shot types with images at minimal acceptable sizes. I can explain the effect of content type on minimal size by the football's XLS, which was different from all other XLS. It depicted small actors on a field that people wanted to be able to see. The music XLS had no actors and the actor in the animation XLS did not move and was hard to see. In the XLS of the news content the people were quite large compared to the football players. The fact that I found significant interactions between content type and shot type at minimal size could stem from other potentially confounding factors. The qualitative feedback suggested an influence of low contrast scenes, text, camera movement and the presence or absence of actors. Considering that across both resolution groups, acceptability at the averaged minimal size was around 66%, service providers would lose a large share of their potential viewers when designing content close to these minimal sizes resulting in viewing ratios of 14 and higher.

11.6.3 Minimal angular resolution

The *minimal angular resolution* depended neither on *content* nor on *shot types*. The effect of *resolution* on minimal angular resolution could be due to a ceiling effect - the 168x126 group could not select larger sizes with correspondingly lower angular resolution than were available. The theoretical minimum at the largest size for 168x126 was 16ppd (11ppd for the 120x90 group). This is supported by the lower standard deviation of the average maximum size (4mm *vs.* 7.5mm) and that the acceptability obtained from the polynomial trend line of the average maximal acceptable size depicted in Figure 55 (84%) is much larger than the values of two lower bounds on minimal size and the bound maximum size for 120x90 content (all between 63% and 71%). In terms the lowest acceptable angular resolution, this seems to be the same for all content and shot types around 14ppd. This is close to the 11ppd that Lund (1993) found through the minimal viewing distances for very large projections of video content in darkened rooms.

11.6.4 Favourite size

The *favourite size* depended on the *resolution* of the content. People preferred to watch higher resolution material at larger sizes than lower resolutions. The average *favourite size* of *news* was smaller than that of other content types. In the 168x126 group news (at 21ppd) did not differ but at 120x90 people preferred to watch news at an average size of 30mm (17ppd) in contrast to all other content types (33mm, 15ppd). Most likely this was rooted in perceived quality of text. People made the fewest complaints about text either being '*illegible*' or '*too hard to read*' at the 30*mm* picture size. Similarly, in study 2 smaller depictions of news had received higher acceptability scores than the largest depiction when the text was subject to encoding bitrate constraints. I will discuss this further in Sec. 12.5. At the favourite size of other content types the rendering of text might result in poorer quality than at a slightly smaller size.

Football on the other hand was preferred at significantly larger sizes due to the XLS. It depicted a far away pitch in which actors were only 12 pixels in height in the original footage and the participants complained that "*it's hard to follow the ball*" and "*I want to see the players more clearly*". One participants comment summed it up very well "*It's big enough but you need to move in with the camera*". At the *preferred size* the actors were about 0.7º tall.

11.6.5 Acceptability of visual experience

In terms of *trading off* size and *definition* the acceptability of the video clips increased until the viewing ratio reached 10.6*H*, at which point the angular resolution was 16.5ppd for the 120x90, and about 8.7*H* for the 168x126 (19.4ppd) video clips. From there on, the acceptability declined and complaints about definition rose as angular size increased and angular resolution declined. My participants made comments about the '*high-definition*' at small sizes but did not try to achieve Westerink & Roufs' maximum picture quality of 32ppd. Although angular resolution of 32ppd was possible to attain in both groups the resulting size was considered too small. Apparently, size concerns must have been considered differently for the acceptability ratings of visual experience. However, the computed acceptability maxima were close to the favourite sizes chosen by the participants.

As in study 1 and 2 the limited encoding bitrate could be a confounding factor for comparing the angular resolution and the potential influence of video quality on acceptability. However, both the actual encoding rate (abr) of both the 120x90 (171kbps) and the 168x126 (181kbps) clips were well below their target. But, since the amount of pixels that are encoded for the 168x126 clips (21,168) was roughly twice that of the 120x90 clips, this should still leave room for possible spatio-temporal degradations due to the higher requirements for the resolution. VQM measures that compared the video clips at 168x126 and 120x90 with their 192kbps encoded counterparts the 168x126 base clips had on average received slightly worse scores (4.54) scores than the 120x90 clips (4.71). Whether this small difference in video quality could have a stronger influence on the acceptability at 14H than the much larger differences in angular resolution (22 ppd *vs*. 31ppd) seems questionable. From the qualitative feedback it seemed more likely that the higher resolution in contrast to the small size was the source of discontent. At the same size the higher resolution group complained more about insufficient size than the lower resolution group (see

Figure 57, left). I also received comments, in which people specifically pointed out the high-definition but complained about insufficient size. Apparently, in rating acceptability of the visual experience, they considered the combined effect of image size and resolution. I cannot rule out the possibility that this effect is an artefact of the experimental design in that people were able to choose larger sizes and did not face a forced choice between a higher and a lower resolution picture at the same small size. However, at a VR of 7 Neuman's participants had also preferred lower resolution content (~44ppd) over high resolution content (89ppd). Taken together, this suggests that - if conservative service designers have to design for relatively small screens - it would be counterproductive to present overly high resolution content. Using a lower resolution could result in higher acceptability although it would yield a lower visual quality as predicted through e.g. Barten's SQRI model and Westerink & Roufs' results. This has tremendous implications for service providers, who could save on bandwidth and deliver a better visual experience to customers at the same time.

11.6.6 Limitations

Further analysis of the data showed that only a quarter of the cells in the 4x4x2 design were normally distributed according to a Shapiro-Wilk test. Some authors claim that the F-test is unreliable if there are deviations from normality (Lindman, 1974) while others claim that the F-test is robust (Ferguson & Takane, 2005, pp.261-2). Although this might call the results of the ANOVA into question, I am confident about the main findings since they matched the qualitative feedback and the results of my previous studies.

Chapter 12

Every person takes the limits of their own field of vision for the limits of the world
- A. Schopenhauer

Discussion & summary

This chapter compiles and discusses the findings from all studies presented in this thesis in relation to previous research.

12.1 Viewing distance

My results on PVDs for multimedia consumption on mobile devices (with different addressability – 115ppi [study 1-4] and 200ppi [study 6]) varied between 25cm to 50cm. This is in line with findings of approximately 35cm by Kato *et al.* on a 166ppi device obtained from people either standing or sitting. Lund's projected viewing distance of 53cm for screen sizes approaching zero height was close to the upper bound of the range for mobile devices observed in my studies. In my studies people on average watched their screens from a closer distance – from around 28cm in study 1 – albeit in a context that was not completely passive but included rating acceptability on the display with a stylus – and 32cm in study 6. The viewing distances of 28cm on a 115ppi device (study 1), 35cm on a 166ppi device in Kato's and 32 on a 200ppi device (study 6) do not support a correlation between PVDs and display addressability. Previous research had identified screen size as the major factor in lab studies that determines PVD for TV (*cf.* Figure 58) but taking into account both Kato's results and my measurements in studies 1 and 6, it appears that people do not adapt viewing distances to mobile devices to increase the angular size of the picture but that the viewing distance depends on the posture the person is taking. Overall this suggests that the PVD does not depend on screen size in solitary viewing on mobile devices. The weight and ergonomic factors of the device should have some influence on the viewing distance but in my studies none of the participants remarked upon them.

12.2 Viewing ratio and size

As presented in the background chapter screen size is a strong determinant for their PVD when people can freely adjust it. People might not adjust the viewing distance on mobile devices to increase the angular size of the picture but if increasing the size of the video window were possible through other means this could make a big difference for the visual experience. Study 6 showed that participants' preferences for watching low resolution content depended first on size. All content types received poor ratings when presented at a viewing ratio larger than 14H. The preferred picture sizes were 32.6*mm* (VR=9.8) and 37*mm* (VR=8.6) for 120x90 and 168x126 resolution content respectively with acceptability between 80% and 90%. The acceptability of video started declining rapidly for viewing ratios larger than 11*H*. On average people required at the very minimum a viewing ratio of 16.3 or 120x90 and 13.9 for 168x126 resolution content – but the average acceptability of these was only about 66%. To put it simply, when given the choice people seek to attain viewing ratios on mobile devices that are close to those in average living room settings (around 8.5H) even for sub-TV resolution content. Further evidence to support this comes from study 3 on the train, in which larger sizes were more acceptable than in the lab but the acceptability dropped off sooner on the train (for VR>8) than in the lab (VR >9.6). In the lab, content shown at 9.6H was just as acceptable as the larger depictions of 6.8H or 7.8H but on the train 9.6H was significantly less acceptable in comparison to those VRs.

With viewing distances considered fixed the only way to adjust viewing ratios to people's preferences is to change the size of the picture. Conservative service providers of mobile TV should deliver content at QCIF resolution (see Sec. 12.3) as a minimum and match the resolution with screen heights of 4cm and larger. Observations from industry confirm my findings and recommendations on size. According to Strategy Analytics (2006), Samsung stated that displays of their first mobile TV phones (33mm in height; a VR of 10.6 at 35cm) were '*probably too small'*, and Nokia and Telia Sonera found that usage rates almost doubled with a screen diagonal larger than 7.6cm (a VR of 7.6 at 35cm viewing distance). Upscaling content at the cost of reduced angular resolution represents an easy way to improve the visual experience, which I will discuss along with its limits in Sec. 12.4.

My participants preferred to watch low-resolution content at viewing ratios that were much larger in picture size than the ITU recommendations for evaluating video quality implied in these settings. I have plotted the ITU's recommended values in Figure 58 along with the proposed preferred viewing ratios based on my results *on preferred viewing size* from study 6 and the results of Lund, Nathan & Anderson, Jesty, Tanton, Kato *et al.* and Ardito of PVD.

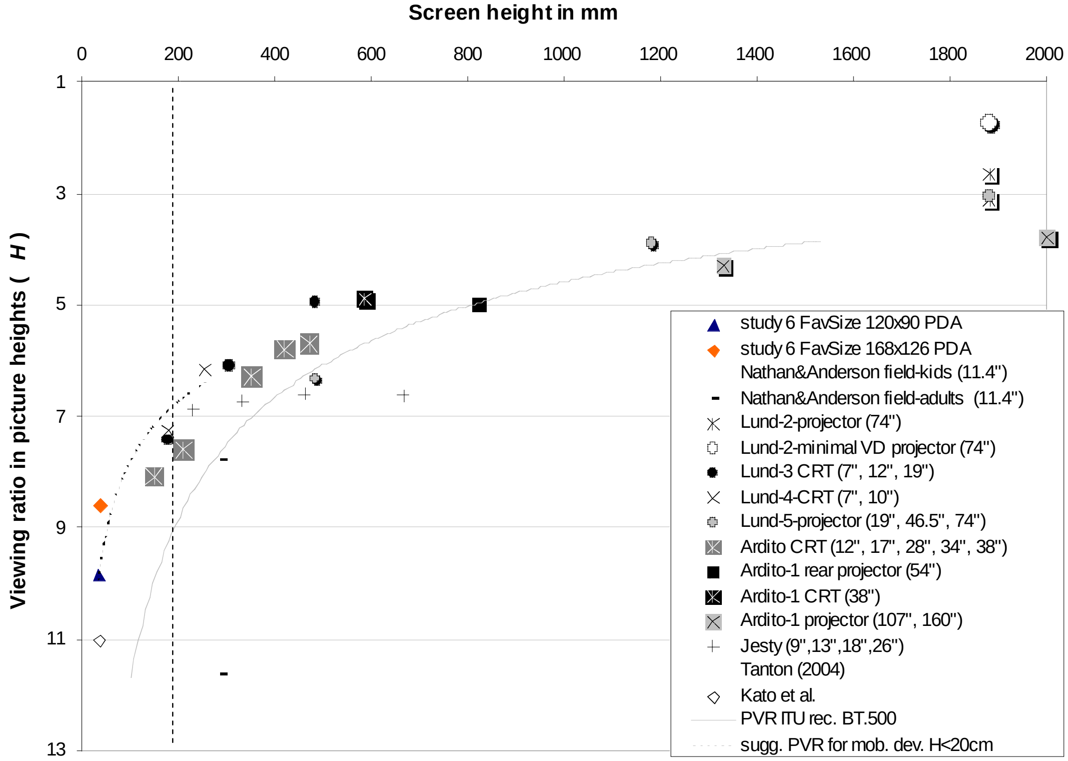


Figure 58: VRs in relation to screen size. All based on preferred viewing distances (PVD) apart from my results on preferred viewing sizes (PVS)

I have included my own suggestion for a PVR for mobile devices (dashed black line in Figure 58) with screen heights smaller than 20cm. The trend line $y=18.152x^{-0.1754}$ that maps the screen height in millimetre to the preferred viewing ratios based on Kato *et al.*, Lund and my values, all of which used QVGA or lower resolution content.

Note that in the field people might have slightly larger preferred sizes as suggested by the results of study 3. Ribchester criticized Jesty's results on PVDs as potentially due to conditioning to typical viewing ratios in the home. People might try to attain similar sizes on mobile devices than they are used to in living room setups. With the advent of larger and HDTV screens this might change and call for even bigger viewing ratios than were preferred by participants in my experiments. Since viewing distances for mobile devices will be more or less fixed for private consumption and people prefer widescreen depictions it makes more sense to express the viewing ratio in terms of picture widths as fractions of the viewing distance, e.g. 0.2*D* would indicate that the picture width was a fifth of the viewing distance.

12.3 Resolution and Adressability

People did not change their viewing distance depending on size or resolution in studies 1 and 6. Since the preferred viewing ratios on mobile devices are similar to living room TV setups I can now look at which resolution people require for mobile TV. In analogue TV resolution was generally synonymous with addressability, i.e. the number of lines of the TV. Within the limits of their addressability digital screens can theoretically present content of arbitrary resolution. I used devices with an addressability of 115ppi and 200ppi in my studies. As detailed in Sec. 2.4.1, high spatial resolution objects are degraded by a reduction of resolution. The two classes of objects that were affected most in my studies and degraded the acceptability significantly were text and players in XLS, each of which will be discussed in separate sections below. Despite these two resolution sensitive items the overall video quality is affected by resolution as shown by Westerink & Roufs. Since the 120x90 content in study 6 only reached a maximum of 80% acceptability at the favourite size I can assume that it was too low to satisfy the entire market. Conservative service providers should deliver content at least at QCIF resolution, which should be acceptable to more than 93% of users when coupled with a screen height of 4cm and larger and an encoding comparable to the one used in this study (around 200kbps, WMV8, 32kpbs audio). By definition this presents the low end of the levels of quality that will be acceptable. Below QCIF resolution could be used for special services with low quality profiles as a large number of participants still found TV content with a 120x90 resolution acceptable given it was presented at an adequate size (9.8H).

The resolution of the content does not satisfactorily explain the results of my studies even when viewing ratios are controlled for and encoding bitrate is fixed. Figure 59 collates the acceptability results of 192kbps clips of study 1, 2, 3 and 6 as a function of their viewing ratios. To remove confounding contributions of shot types I weighted the acceptability results of shot types in study 6 with their relative occurrence in the footage used in studies 1, 2 and 3. Although the results for study 1-3 followed the same trend as the results from study 6 the reduced acceptability of the clips on the 115ppi device was remarkable, especially since the clips in study 1 and 2 were presented in the same laboratory environment as in study 6. The discrepancy could stem from the differences in the displays resolution (addressability), contrast, stimulus structure, rating method (rating what is available *vs.* freedom to adjust to preference) and the fact that viewing distance was controlled for more precisely in study 1 than in the previous studies.

Even if the contrast was lower on one device than on the other it should not make a big difference since the contrast sensitivity is almost at its maximum at a spatial resolution of around 21ppd. Therefore, differences in contrast should not account for the different levels of acceptability at 21ppd in Figure 60. Of the remaining possibilities the stimulus structure and device addressability are the most plausible explanations. In study 1, 2 and 3 the clips had been presented as part of a 2:20 minutes clip that was degrading in quality over time whereas in study 6 they were only shown for up to ten seconds. In study 1 and 2 the acceptability of both football and news dropped disproportionately after the first segment of 20 seconds at 224kbps. However, the drop offs observed were not large enough to explain the differences in the comparison at hand.

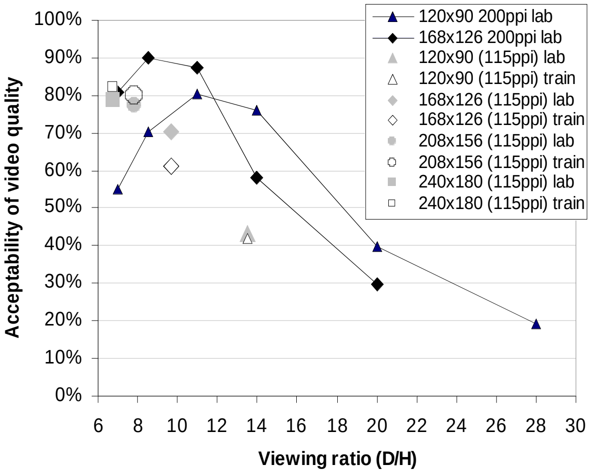


Figure 59: Acceptability at 192kbps in study 6 (solid diamond and triangle)**, study 3** (white)**, and combined results of study 1 and 2** (grey) **by VR**

In (McCarthy *et al.* 2004b) QVGA football content encoded at high bitrates of 448kbps, shown natively on a 115ppi device at 5H (21ppd) with players of 25 pixels in height was only acceptable to 83% of the participants. This can be partly attributed to the relatively large angular screen size with an inadequate angular resolution (5H, 21ppd) – the results from study 6 suggest that people prefer to watch this content at a higher angular resolution because QCIF resolution content was already watched at 21ppd and the tendency would suggest a higher preferred angular resolution (*cf.* Figure 61). But the results of the comparable XLS shot in study 5c resulted in a high acceptability (about 95%) with a lower angular resolution (19ppd) at 8.5H. So neither overall size, actor size, encoding bitrate nor angular resolution of the content offer compelling justification for the lower acceptability of football content as observed in (McCarthy *et al.* 2004b).

In study 1 the participants watched the clips from 28cm on a 115ppi display resulting in an angular resolution of the display of 21ppd whereas in study 6 they watched a 200ppi display from 32cm (45ppd). Higher addressability should result in fewer artefacts such as aliasing.

This leaves device addressability as the

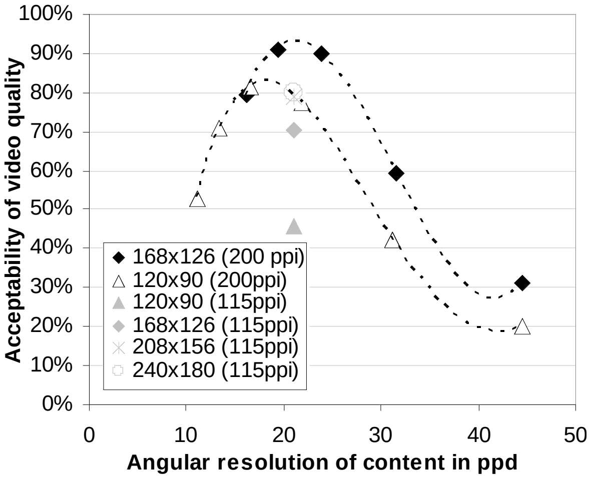


Figure 60: Acceptability depending on angular resolution

likeliest source of the differing results in study 1 and 6 as shown in Figure 60 and Figure 59. Should this hold true it would suggest that future displays with even higher resolution might still improve portraying video content non-natively. Comparisons with the higher resolution clips 208x156 and 240x180 from the previous lab study are harder to make, as I cannot know the exact resolution of the content due to the spatio-temporal compression of 192kbps. This is a general limitation of my results on angular resolution requirements – apart from the news study in which text was shown in its original broadcast quality (see Sec 12.5).

12.4 Trading off angular resolution for size

The range of viewing ratios and angular resolution at which people find it acceptable to follow video content is large. Both size and the available resolution of the content must be taken into account for the best presentation of TV material on mobile devices. Qualitative feedback about insufficient definition in study 6 showed that the video quality diminished once the angular resolution dropped below 32ppd matching results from Westerink & Roufs but the acceptability of the visual experience dropped only once the angular resolution dropped below 21ppd for 168x126 (8.6*H*) and 18ppd for 120x90 (9.8H). The difference in preferred viewing sizes depending on resolution was large given that

1) Lund suggested to seat viewers of HDTV content 15% closer to the screen than for SDTV at the same height but with half the resolution and
2) video quality assessment of HDTV and SDTV as in ITU's BT.500-11 do not differ in the recommended PVR.

Content shown on mobile devices at larger size was generally more acceptable than smaller sizes at identical encoding bitrates unless text was involved. The acceptability of QCIF content with small text declined once the angular resolution dropped below 21ppd.

In study 6 I found that QCIF resolution content, if not shown at sufficient size, might yield lower acceptability than content of lower resolution at the same size. Although surprising, this is not the first occurrence of this mismatch phenomenon. Neuman's (1988) participants preferred lower resolution (~44ppd) over high resolution TV content (89ppd) at 7H.

Westerink & Roufs suggested that people would choose their viewing distances to attain the best subjective quality – an angular resolution of 32ppd (16cpd) – irrespective of picture width. My results in the domain of mobile devices did not support this but showed that participants' preferences for watching low resolution content depended mainly on size (see Figure 60). In study 6 acceptability peaked at 16ppd (10.3*H*) for 120x90 and at 20ppd (9*H*) for 168x126 resolution content. For higher resolution content Westerink & Roufs' projection might hold but the strong influence of size on the preferred viewing ratio as apparent in Figure 58 does not support their assumption. This is not to say that the perceived video quality is not at a maximum at 32ppd. But the overall visual experience that people are trying to achieve especially in terms of size takes precedence over maximizing the perceived video quality. However, in my experiments the PVS of low resolution content (QCIF and below) resulted in a low angular resolution and people avoided angular resolution close to and below 12ppd. In one of Lund's experiments participants chose their PVD of low resolution content (220 TV scan lines, marked as Lund-5 in Figure 61). For the largest depiction (1.88*m* in height) they preferred a viewing ratio of 3.1, which resulted in an angular resolution of 12ppd. In another experiment (Lund-2) he assessed the minimal viewing distance, at which the participants were willing to watch projected SD- and HDTV content in a dark room. For SDTV content the participants chose a viewing distance of 1.7*H* resulting in an angular resolution of 11ppd. Together with Lund's results from large projections of TV content in dark rooms it seems likely that 12ppd will hold as a lower bound for angular resolution and will thereby limit the achievable viewing ratios by scaling low resolution mobile content.

In Figure 61, I have collated the preferred (PVD) and minimal viewing distances from the aforementioned studies by Lund, Ardito, Ardito *et al.* and Nathan *et al.* and plotted them in terms of the resulting angular size and resolution. Results obtained in dark rooms are marked with shadows. The assumed lower limit in terms of angular resolution is marked with a dashed/dotted black line. I added the results of my study that were based on preferred viewing sizes (PVS) rather than PVD (except for Lund-2, which was based on minimal acceptable viewing distance).

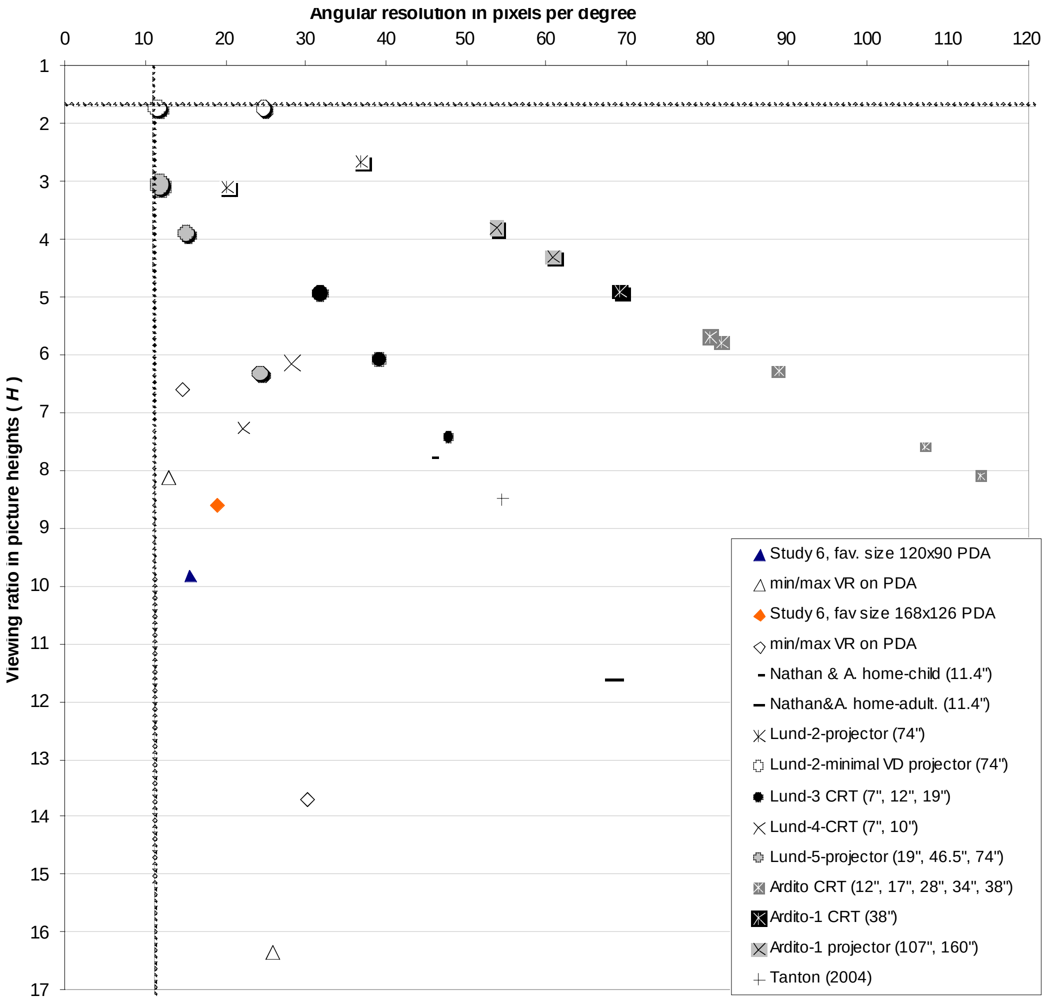


Figure 61: Comparison of results obtained by Lund, Ardito *et al.* and Nathan *et al.* on preferred viewing ratio with my results on preferred size

Large viewing ratios (below 2H) as observed by Lund and Ardito are not a current concern for sub-SDTV resolution content presented on mobile devices even as large as the iPad due to the limiting angular resolution. But with this lower bound established across different sizes I can revisit the results of Lund's study. He did not mention why people would not move in closer than 1.8*H* (25ppd) for HDTV resolution content. Since the angular resolution was not at the lower limit yet another explanation is required. Fujio's research into HDTV showed that watching dynamic content at closer than 4H may induce fatigue and motion sickness. Therefore, a possible explanation for Lund's finding could be that at these small viewing ratios people have to make increasingly more use of head movements to be able to watch everything on the screen. The horizontal visual angle was 42º horizontally, 22º upward and 10º downward without tilting the head upwards. According to Hatada *et al.* people can only cover 30º horizontally and 8º upward and 30º downward by eye-movement alone.

12.5 Text

Text is a major concern for content producers. Most text in TV production is not sampled by a camera but is digitally inserted and therefore not reduced as much in resolution by interlacing and sampling as camera captured material. On interlaced displays moving text further loses some of its resolution and guidelines for the minimum sizes for SD- and HDTV exist. Since people are keen on achieving viewing ratios on mobile devices similar to living room TV depicted text would attain the same angular size but yield a much lower resolution. Based on the results from study 1 I cannot recommend reusing unedited TV news for mobile consumption when targeting QCIF resolution. Text quickly dropped below five pixels in height and drastically reduced the visual experience. This was true for text that was presented in the main window – in the centre of the audience's attention – and in the periphery of the screen. The changes made to the original clips in study 2 ensured text legibility on 115ppi devices, could be automated and reap substantial benefits in acceptability.

In study 1 smaller depictions of news received higher acceptability scores from non-native speakers than the largest depictions. Study 2 showed that this was due to the fact that the encoding bitrates were not high enough. Native speakers, however, appreciated the larger text, which compensated for lower quality. The same was true for participants on the train. Despite the reduced video quality – as pointed out in the comparison of lab results study 1and 2 - enlarging text from 20.5 (208x156) to 22.5 arc minutes (240x180) increased the overall acceptability of news content on a 115ppi device on the train. ANSI specifies limits of 16 minutes of arc (*cf.* Sec. 2.3.5) and my results showed that people watching on mobile device valued further increases in size and resolution. At the two favourite sizes in study 6, text was 17.5 (17ppd) and 18 arc minutes (19ppd) tall. In study 6 the minimum angular resolution of news (14ppd) was no different from other content types that did not contain text but people liked the text to be crisper – preferably not below 19ppd for QCIF content on a 200ppi device. Recall that in study 6 video content was presented on a high resolution display that resulted in a 45ppd raster but in studies 1-3 the display resulted in a 21ppd raster. Aliasing of text should therefore have been less of a problem in study 6. Lund did not comment on whether text was visible in the content used in his experiments in which participants were willing to watch SDTV and sub-SDTV resolution content at an angular resolution far below 20ppd. From the description of the content it seems that text did not feature prominently in it, which would also explain the low angular resolution (close to 10ppd) at which people were willing to follow the content.

12.6 Shot types

Tailor-made content production in terms of length and visual style is expensive and many service providers want to simply recode broadcast TV content for the appropriate target device resolution to simplify system design. Is this possible given the constraints of mobile services? My results showed that sports content suffered most in terms of acceptability in study 1 at smaller sizes at 21ppd angular resolution. However, at that point it was not clear from the qualitative feedback whether this was due to insufficient size, resolution, encoding bitrate or the addressability of the device. Too many variables were confounded and obtaining the feedback at the end of the session did not provide the necessary discrimination to identify the source of the problem. The follow-up analysis of study 4 showed that the acceptability of football's XLS - depicting players from far away - dropped disproportionately when presented natively at 21ppd, at resolution of 208x156 and below. Although objective PSNR scores suggested that the XLS's image quality was superior to the LS's and VLS's the participants found the XLS less acceptable. To some degree this ruled out encoding bitrate as the possible cause but size and resolution were still confounded. Only studies 5a-c in conjunction with study 6 made it possible to understand that players required a resolution of 15pixels but more importantly they needed to be depicted at a sufficient size. Adaptation of XLS to achieve these two requirements would be valued on mobile devices and I will discuss this and the requirements in more detail in relation to zooming in Sec. 12.7.

Shot types like MCU and MS that portrayed more detail had smaller acceptable size limits than other shot types in study 6. At first sight this should be encouraging for content producers relying more on close shots. However, the preferred size of these shots did not differ from other shot types and therefore this kind of production would only prove beneficial for QCIF content that would be shown on displays with heights smaller than 22*mm* (VR>14). The field results from study 3 and as discussed in Sec. 12.2, however, indicated that this will not be acceptable to a large portion of the audience. At QCIF resolution the minimum required for mobile TV the minimal size requirements MCU and MS were not different in from LS. At sufficient size and resolution for mobile TV there seems to be no need to pay attention to shot types apart from the XLS.

12.7 Zooming

In Sec. 12.3 I argued that mobile TV services require a minimum resolution of about QCIF and a VR of 11 and smaller (in Sec. 12.2). Studies 5a-c showed that this blanket statement might be too general and that sports content - specifically football - employing XLS shots might have further, more stringent requirements. Various complaints in studies 1 and 3 referred to insufficient detail of content. Since football is high motion content this might be partially due to the insufficient encoding bitrates used in my studies. However, study 4 provided some evidence that the video quality itself was not at the heart of this problem but that in some cases the selected resolution was too low and sizes of the player and the ball too small. Depending on whether a living room VR can be provided on mobile devices or not zooming might serve two different purposes. For living room viewing ratios zooming could provide more detail of depicted objects whereas the increased size of objects would be the most valuable trait if the screen size were insufficiently small. Since resolution is the main difference for mobile services at those viewing ratios the main differentiator would be the greater amount of detail of the zoomed-in content. Considering

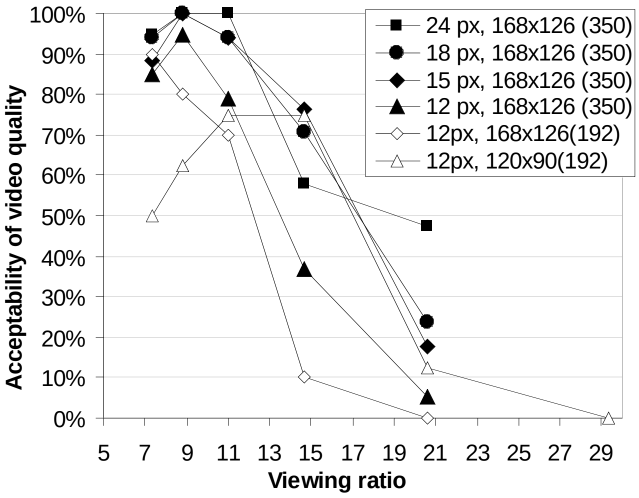


Figure 62: Fans acceptability of XLS in study 5c (solid) and 6 (white) by player height in pixel at 192 and 350kbps

the large range of angular resolutions at which people found viewing acceptable (*cf.* Figure 61) the added value of zooming to increase the amount of detail of objects seems to be limited especially since zooming would need to sacrifice parts of the image.

My results on zooming confirm this line of thought. As with overall image size, player size was more of a concern than resolution (*cf.* Figure 52). Figure 62 collates the acceptability results of football XLS from study 5c (solid) and 6 (white) of fans only. The value of zooming for an acceptable visual experience diminished once the viewing ratio of the whole picture was below 8.5 for content encoded at 350kbps. For viewing ratios of 14 and larger there was still a large benefit of zooming but the overall acceptability was low (75% at best). Viewing ratios between 8 and a maximum of 11 should result in the best experience of QCIF football content. At 11H zooming increased the acceptability of XLS tremendously but started reaching a plateau as the angular size of the players approached 0.8°. To be acceptable to all fans the size of players needed to be between 0.7º and 0.8º regardless of their resolution. At a VR of 8.5 this requirement was typically met and fans found sports content at QCIF resolution encoded at 350kbps even better than acceptable when the resolution of players was higher than 15 pixels. In study 5b football fans had preferred player sizes from 0.5° to 0.7° for VRs of 14 to 8.5 when it meant relinquishing contextual information from the pitch. I can therefore conclude that upscaling QCIF content with a factor of 1.35 from a VR of 11 to 8.5 will improve the visual experience more than zooming into the content at a VR of 11 with a high magnification factor (e.g. 2 as). Zooming will yield the biggest benefits on devices on which VRs of 8 (picture heights of around 4cm) are not possible but an angular size of 0.8º for sports players in XLS can be achieved. Fans might appreciate zooms to be optional to maximize the view of the pitch. Other sports, for example ice-hockey, might have different requirements.

Methods to increase the size of certain objects, e.g. the ball in sports content as in (Nemethova *et al.* 2004) had two problems. First, they only worked well in the authors' studies for very low encoding bitrates, which were acceptable to a minority of participants throughout my studies. Second, because people want to watch content at living room viewing ratios, increasing the size of the object further does not make much sense. Since the size of players was also important to my participants zooming into XLS boosted acceptability when living room VRs could not be attained by increasing the size of both the ball and the players.

All shot types of other content types did not differ in the preferred size in study 6. Although this might be partly due to people's preferred trade-off between size and angular resolution I found for sports XLS that people did have some margin within which they traded off resolution for size or in the case of news size for resolution to improve text quality. Since the preferred sizes of other shot types did not differ the idea of zooming seems to have little application outside of sports XLS unless content needs to be adapted to very small screen sizes, which resulted in unacceptable visual experiences for a majority of participants.

12.8 Encoding bitrate

The required encoding bitrate depended to a large degree on the content. Unsurprisingly, low motion animation was the least demanding in terms of encoding bitrate, while camera panned football pitches were on the other extreme of the spectrum (*cf.* Figure 25, p. 105). Reducing the resolution of the encoded video to achieve higher encoding bitrates per pixel as envisioned in Sec. 4.6.3 made only a difference at very low encoding bitrates and for text for the video dimensions selected in studies 1 to 3. Larger depictions with higher nominal resolution at the same encoding bitrate were generally more acceptable due to the large contribution of size. Figure 63 includes dashed trend lines for the encoding bitrates used in study 3. For all encoding bitrates of 96kbps and higher a target resolution of 240x180 achieved the highest acceptability. The only content type for which I observed a possibility of increasing acceptability by reducing the encoding resolution was news. Text that featured prominently in news content is a high frequency visual part, which benefited from gains in encoding bitrate despite the smaller overall sizes in study 1 and 2 in the lab.

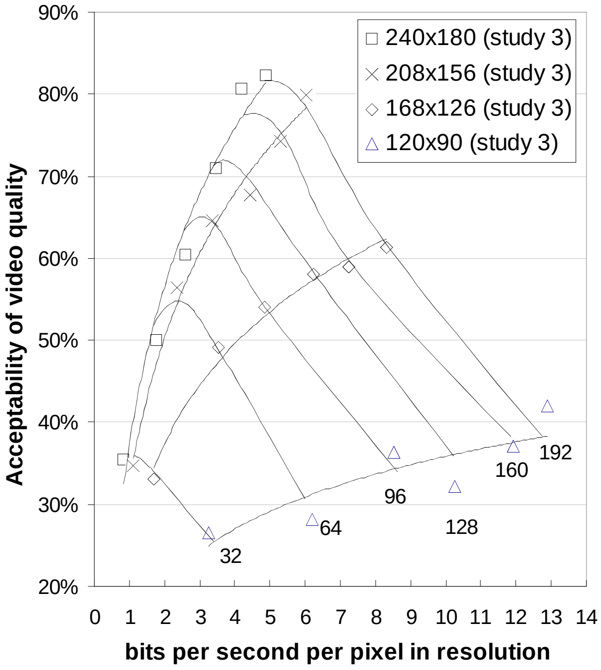


Figure 63: Trading off resolution and size for encoding bitrates on the train

provides a cross study comparison of football content between study 1 with (McCarthy *et al.* 2004b), which used a nominal QVGA resolution. Both used the same device for the depiction of the videos a iPaq with a 115ppi screen. I included finely dotted trend lines in , each holding size and resolution constant and manually fitted bold trend lines holding encoding bitrates constant (192kbps in grey, 160kbps dotted, 128kbps dot-dashed and 64kbps dotted). Increasing the resolution from 240x180 to QVGA resulted in lower acceptability when football was shown natively on the screen. For all other depictions lowering the resolution/size below 240x180 resulted in lower acceptability. However, this is only representative for a native depiction on a 115ppi device and up-scaled content of e.g. QCIF resolution can result in equally acceptable depiction on a 200ppi screen as seen in study 6. Note that the procedure and the content differed between (McCarthy *et al.* 2004b) and study 1 but both depicted football content.

At the beginning of my research the versions of WMV used in my experiments was reasonably new but in the meantime encoders have become more efficient and fewer bits per second per pixel resolution will be required to reach the same visual quality. The large contribution of size to the visual experience, however,

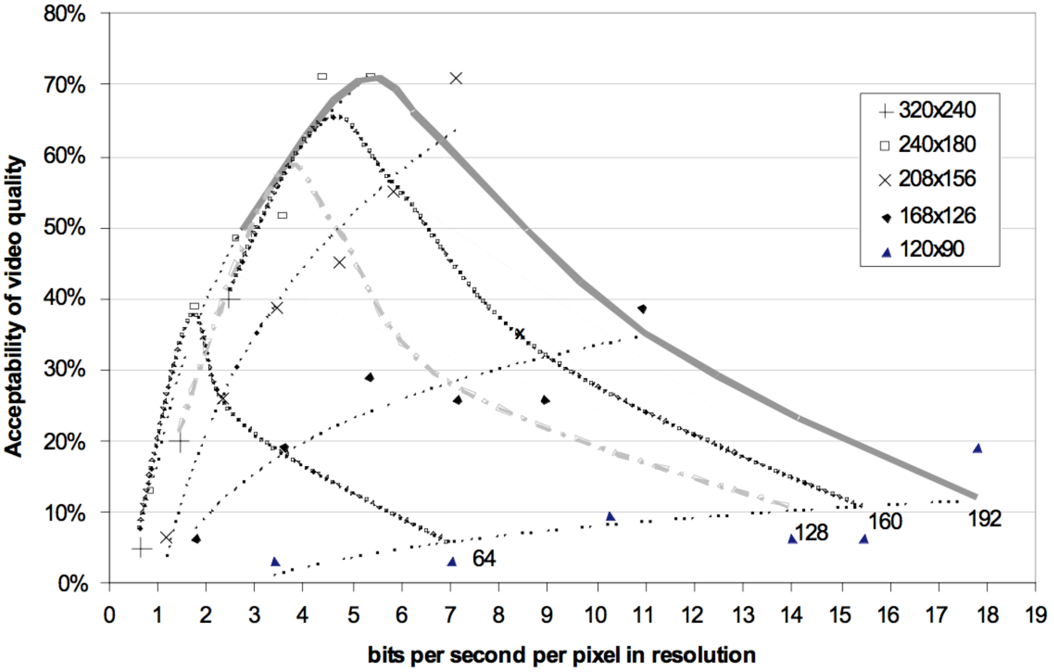


Figure 64: Cross-comparison of football content in (McCarthy et al. 2004b) and study 1

should remain. In order to achieve higher qualities (beyond the acceptable level) the encoding bitrate budgets might be more important than in my studies.

12.9 Models of multimedia and experience

As presented in Sec. 3.8 most multimedia models follow an additive or multiplicative approach. Better quality in one medium is assumed to improve overall multimedia quality. This assumption appeals intuitively but runs against some of the research findings presented in Chapter 2. This thesis does not provide a model for multimedia quality and it looked at visual experience rather than video quality. But the evidence provided suggested that acceptability might not be properly described by a purely additive or multiplicative approach. Study 1 showed that video clips displayed on mobile devices were more acceptable to participants across all video encoding bitrates when it was supported by lower (16kbps) than higher audio quality (32kbps). Non-native speakers found news footage in study 2 less acceptable when it was accompanied by continuously high quality text than by text that degraded with the video. In study 6 participants assessed higher resolution clips less acceptable than lower resolution clips when they were shown at small sizes. All three findings suggest punitive ratings in cases where one media dimension exceeds other dimensions in fidelity. One way to explain these phenomena is that high quality in one medium makes shortcomings in other media more obvious or creates higher expectations for it. This can be likened to Feistinger's cognitive dissonance theory in social psychology, which holds that a person's perception of logical inconsistencies can induce negative emotional states, which can then lead to worse evaluations.

According to my findings and literature research models of audio-visual experiences would have to include:

- a notion of viewing ratios,

- aspect ratios,
- an adjustment for deviations from preferred viewing ratio and angular resolution,
- a measure of text legibility and quality,
- adjustments for discrepancies between audio, video and text quality and
- an adjustment for small player sizes in field sports at large VRs.

12.10 The train *vs*. the lab

In studies 1 and 2 the depiction of news with text benefited from gains in encoding bitrate despite the smaller overall sizes in the lab, but on the train the gain in textual rendering was outweighed by the reduced size and the largest size became the most acceptable depiction. Although my research showed that lab experiments might provide conservative estimates of acceptability in the field, this was not true for all observed factors. For effects that were not fully understood yet, e.g. image and text size, tests in the field provided more insights and were advisable to validate, possibly correct and enhance laboratory results.

12.11 Video quality - a poor proxy for visual experience

Research on video quality often assumes that people prefer to view video at the highest possible quality. Westerink & Roufs suggested that people would choose their viewing distances in order to attain the best subjective image quality – an angular resolution of 16 cycles per degree (32ppd). This approach was based on people's ratings of pictures of different sizes and resolution at different viewing distances. They were not asked to choose their PVD. The results presented in this thesis show that participants' preferences for watching low-resolution content on mobile devices depends mainly on size – depending on the content's resolution they preferred viewing ratios between 8.5 and 10, which resulted in an angular resolution between 19ppd and 15ppd. From the complaints about insufficient resolution in study 6, I learned that angular resolution became a concern only once the picture was big enough - a VR of at least 14 or smaller. The acceptability of QCIF content dropped off when angular resolution declined below 20ppd. Between Westerink & Roufs' optimal visual quality of 32ppd and this 20ppd threshold the acceptability of QCIF content presented on mobile devices improved by trading off angular resolution for larger size.

The participants in my studies preferred to watch low-resolution content at viewing ratios that yielded larger pictures sizes than the ITU recommendations for evaluating video quality in these settings. Recommendation BT.500 *implies* a PVD for small screen sizes in the range tested here of at least 15H. Considering the tremendous difference in acceptability for the different viewing ratios in study 6 (*cf.* Figure 55 on page 144) and the additional value of size while watching video content on the train in study 3 video quality measured under ITU recommended would poorly predict the visual experience. Recall that Westerink & Roufs' results showed that video quality increased with angular resolution for resolution below 32ppd (*cf.* Figure 61). The small depictions would yield high angular resolutions and result in favourable video quality ratings. Participants in study 6 pointed out that small depictions had high definition but their experience was not acceptable due to the small size.

12.12 Acceptability

The label 'acceptable' and the context in which it was placed in the question "*do you find the video quality acceptable for a mobile TV service*" were intended to elicit responses that included the experience as whole. The fact that the maximum values of acceptability coincided with people's preferences in study 6 is a desirable property of an experiential measure. Obviously this label operates at the threshold of what visual experience people find acceptable and premium services might be interested in offering better experiences. This could be achieved by using other labels such as *excellent* or *very good* when it is framed within the overall experience of the service. Their alignment with users' preferences would need to be validated, however, through preference ratings for example. This would prevent previous operational mismatches of e.g. optimal viewing distances based on Westerink & Roufs' video quality, which did not coincide with people PVDs.

Rating video quality with a stylus on the display in studies 1, 2 and 3 might have caused people to hold the device closer than they normally would when viewing content on a mobile device. This might have resulted in a viewing ratio slightly smaller than in study 6 in which people were able to provide their feedback verbally. However, the method in the former studies did not require interaction with an experimenter. This might have altered participants' ratings because involuntary verbal and non-verbal cues from the experimenter might have prompted the participants to adjust their ratings to please the experimenter. So one has to be careful that the way the ratings are collected does not alter what people are supposed to rate. Sitting slightly set back from the participant should have reduced non-verbal cues, however.

Ideally, researchers would only alter a single factor between experiments to allow for better comparison between studies. However, this approach might not be possible because of technical restrictions. In my case the higher resolution PDA also had higher contrast and luminance than the lower resolution PDA. At the same time using different setups and different ways of collecting the same type of rating in a set of studies increases the external validity of the results - a desirable property of research.

12.13 Between intimacy and immersion

Mobile device screens can be easily shared by their users with a few other people as long as the resulting viewing ratios do not become too small (see Sec. 12.2). Participants in my experiments had different favourite viewing conditions. If sharing the screen is the defining criterion of mobile TV viewing as suggested by Harper *et al.* (2008), devices offering people the possibility to adjust viewing sizes should help in this process of letting others co-view content, Their *watching-to-show* experience might support a different kind of social TV viewing that focuses on "*being together*" as described by Taylor *et al.* (2002) but includes intimate spatial proximity. This has proven popular for parents when watching content with children (Södergård 2003) on mobile devices. Watching content on a mobile device in a group requires initiation by its user and involves a gesture that carries more social meaning than if a living room TV were turned on and watched by multiple people.

With mobile devices people will be able to moderate their participation from being completely committed to what a group is watching on a shared TV, to previewing different content on a mobile device and to just share a common space and be immersed mostly in the content on a mobile device. Headphones will provide a better audio separation and enable a person to immerse himself in the content and disengage from others, especially in conjunction with a personal screen (O'Hara *et al.* 2007).

12.14 Limitations

There are a number of potential limitations to the research presented in this thesis. Most notable is the young age of the participants, whose eye sight was very good when compared to that of an older population. Since I was testing a visual medium, this does not allow me to know how vision-impaired viewers would experience mobile TV. Viewing distance or the required use of near-sight glasses could pose a problem in mobile contexts, especially if people need to quickly change from near sight to far sight for example when switching from the screen to the environment – but this poses a problem for mobile devices in general. Participants with less than 100% visual acuity were in general more forgiving of low video quality than people with 20/20 vision. As a new technology the appeal to older people might not be an initial concern but only come up once the technology is widely adopted. At that point services usually perform better as economies of scale are leveraged.

Using only parts of the screen in my studies could have prompted participants to judge the acceptability of video in terms of the relative size of the picture to what the device could display. They might have penalized unused screen estate and on a mobile device that used all of its pixels for the display of video the perceived quality might be higher. However, the results I obtained in studies 1 through 6 compared favourably to the results in (McCarthy *et al.* 2004b) in which content was displayed in full-screen mode on both a 3G phone and the same PDA used in studies 1-5.

My studies excluded visual artefacts due to imperfect reception and other network transport errors. These distortions might influence people's preferences in terms of e.g. size. New encoders might interact with the network error profile and make more efficient use of encoding bitrates. Screens with higher resolution, contrast and luminance might further affect people's choices as seen in the comparison of acceptability between study 1 and 6 and I did not control for the contrast and luminance profile of the content. Both affect the modulation transfer function and therefore confound perceivable resolution.

The selection of content used in my studies could affect the external validity of my results. I concentrated my studies on content types that I had identified as attractive to mobile TV viewers and used samples from TV and DVD so the content was both relevant and realistic. Other samples of content and shot types could yield different results but given the quite diverse content types this seems unlikely. However, my most important findings on preferred viewing conditions and minimum angular resolution should generalize for both practical and research purposes. The minimum acceptable angular resolution (~14ppd), for example, can be considered a general threshold for acceptability as it depended neither on content nor on shot type. The effect of resolution on it was most likely due to a ceiling effect and is backed up by similar values from Lund's research. Apart from small adjustments for small text and small actors my main finding about favourite sizes depended primarily on resolution.

I used relatively low resolution and encoding bitrates in my experiments to find the threshold for an acceptable mobile TV service. This might seem overly pessimistic in terms of what mobile TV service providers might be able to offer. Higher resolution content would surely improve the visual experience. Both DVB-H and DVB-T enabled mobile devices target higher resolutions and encoding bitrates through broadcast. But wireless spectrum is limited and offering more content to people might take precedence over resolution of content. Even for these services low resolution might play a role to reduce channel switching times without sacrificing the bandwidth savings of few key frames. Some hybrid solutions target channel switching augmented by data delivered through 3G or higher mobile networks (Hsu & Hefeeda 2009). Unicast delivered resolution material could fill in until the next key frame is broadcast and thereby increase the user experience in terms of channel switching. Furthermore my results on minimal angular resolution are independent of the resolution delivered in a given service but can serve as guidelines for any service employing video material.

Chapter 13

Conclusions & future trends

In a time of drastic change it is the learners who inherit the future. The learned usually find themselves equipped to live in a world that no longer exists.
- Eric Hoffer

In this chapter I revisit the research goals from Chapter 1 in light of my findings as discussed in the previous chapter and maps out the conclusions for different target audiences such as service designer, content producers, researchers of video quality and the emerging QoE research community.

Substantive findings include preferred mobile device viewing distances and ratios, the influence of size, text legibility on acceptability of a mobile TV service, the different effects of size on the different shot types, optimal zoom factors and the limits of zooming into content.

Methodological findings include the use of qualitative feedback, complaining aloud, triangulation of quantitative results through frequency counts of qualitative complaints and the strength of a mixed methods approach.

For the sake of convenience the research goal is repeated below.

1. Which factors affect the QoE of mobile multimedia services specifically mobile TV, and how exactly?

The supplementary research questions were:

2. Which of the existing methods are suitable for establishing and measuring these factors?
3. Under which conditions does content ported from TV to mobile devices result in a satisfactory visual experience?

The following section summarizes my findings for service designers and content producers 13.1.1, for researchers of video quality (13.1.2) and for the emergent field of QoE (13.1.5). For more detailed conclusions I refer the reader to the sections on substantive and methodological contributions within this chapter.

13.1 Conclusions for different target groups

13.1.1 Service designers and content producers

Based on the synthesis of previous research and my own findings I can provide the following recommendations for service designers and content producers. For service designers the most important concern in designing mobile TV services should be ensuring that the service is delivered to devices with sufficient screen size (on 4:3 screens a minimum of 4.5cm in height). Repurposed TV content should be delivered at a minimum of around QCIF resolution. For consumption in indoor settings the most popular angular resolution of QCIF content on a high-resolution display (200ppi) is around 20ppd, (approximately 4cm picture height). For larger depictions users prefer an angular resolution beyond 20ppd. This would mean delivering content with a resolution that is more than proportionally higher than the increase in picture size. As a lower limit for frame rate 12.5fps should be acceptable across entertainment content types. Channel change times should be minimized, ideally to less than a second. The service designers need to evaluate the trade-off between decreased channel switches and incurred increases in terms of the encoding budget with a given encoding format.

Service designers should require content producers to adjust text in content to the lower resolution capabilities of mobile devices. Adaptations in terms of used shot types should not be necessary for mobile TV content with the exception of sports content like football. XLS should ensure a minimum size of players of one degree. For this kind of content an optional zoom could be triggered by the user and either focus on the area of interest or allow for user controlled panning.

13.1.2 Handset hardware manufacturers

Handset manufacturers should ensure a minimum screen height of 4.5cm. Wide screen displays should improve the visual experience. However, the overall user experience that includes user concerns in terms of device portability and the overall form factor might impose limits on screen width. Battery life was of the utmost concern for participants in my studies. Watching TV on mobile should not compromise or infringe on the most important use of the device – staying in touch with other people. Users will need to feel confident about their ability to gauge remaining battery life.

Recent advances in display addressability, such as Apple’s retina display, should prove valuable in displaying content non-natively. My research cannot answer how much gain in visual experience a perfect canvas in which pixels become invisible from typical viewing distances will provide. However, cross comparisons between my studies on 116ppi displays and 200ppi displays even for an acceptable visual experience were pronounced and suggest potential gains especially when higher fidelity than acceptable is sought. Mobile displays with wide viewing angles should prove popular by supporting a *watching to show* user experience in which the user shares the screen with a small number of co-viewers.

13.1.3 Handset software manufacturers

In terms of the interaction with the content pause and mute functions are essential for use on the move (Knoche & McCarthy 2005). Software manufacturers can help reduce the perceived delay by, for example, displaying pre-cached clips such as advertising in accordance with service designers.

For football and similar content a user controllable zoom that allows for more detail either by displaying only the standard safe zone or an area of interest with user controlled panning should improve the user experience. Spatial resolution can be greatly reduced after a channel switch up to a half second. For services with limited bandwidth budgets, separate delivery of textual content should significantly improve user satisfaction while reducing the required encoding bitrates.

13.1.4 Video quality research

Research in video quality should employ viewing distances that are indicative of actual use. For mobile video quality research mobile devices should be used at representative viewing distances between 30 and 40cm and definitely within a range of 20 to 50cm. The size at which video stimuli are presented should reflect people’s preferred viewing sizes. Although the currently suggested viewing ratios of the ITU recommendations on TV pictures reflect people’s preferred viewing ratios under lab conditions, these are not representative of actual home use. If research showed that typical VRs do not have an impact on video quality assessment ratings the current approach would be justified. However, so far this has been shown only for VRs between 2 and 4. This is especially problematic for mobile multimedia consumption, which occurs at higher VRs.

Of greater importance is the conclusion that video quality researchers should abstain from equating video quality with QoE or visual experience. My research showed the limitations of standard (PSNR) and leading objective quality measures (VQM) to approximate QoE. Despite its important role for visual experience, video quality is only one of a number of factors that people need to consider to maximize their visual experience. Ideally, video quality research would embrace the bigger challenge of VX and QoE by including such factors as viewing distance, VR, aspect ratio, addressability, frame rates, shot types, text and content types, to name but a few. This increased scope will require a corresponding extension in the methodological repertoire. Qualitative feedback should be gathered from video quality assessors to better understand on which aspects they base their ratings. When categorical rating scales such as the ITU's ACR are used the experimental design should ensure that the required transformations of the ratings be applied to turn them into a linear scale and allow for averaging of scores. To improve benchmarking and the comparison of different codecs video quality publications should include the encoding bitrate budgets (in bits per pixel per second) at which ratings were obtained.

13.1.5 QoE research

At its current stage, QoE research should focus on identifying further QoE dimensions and finding methods and metrics to measure these reliably. For example QoE research needs to explain why, to what extent, and under what conditions people prefer wider aspect ratios, larger sizes, increased contrast and 3-D displays when optimizing their multimedia experiences. This optimization will require trading off other dimensions and the value of the experience needs to be understood in order to inform the design of the involved devices, services and applications. Research of QoE should employ a mixed methods approach to ensure that participants of studies actually assess the parameters under study and do not rely on other potentially hidden variables. Any QoE focused research should rely on participants that are interested in the content or services and employ stimuli of sufficient length. The results of my first study suggested that the extent of the detrimental effect of small sizes on the visual experience became apparent only after people had watched content for more than 20 seconds.

Qualitative methods should be used to elicit, which factors contribute to QoE and how. Furthermore, qualitative methods can help to disambiguate contributions of confounded parameters and identify new parameters that need to be considered. Feedback should be elicited in a setting that resembles as closely as possible the actual experience. Ideally, feedback would be provided while the participants experience the service or application. This avoids the problems encountered during study 1 in which attribution was difficult because too many factors were confounded in the debrief interviews. To allow for intra-stimulus feedback, the measurements of QoE cannot get in the way of the experience itself. This favours unobtrusive and low-effort methods. In the later stages of QoE research identified dimensions can be rigorously tested with quantitative methods to assess their contribution and to aide model building in this new domain. Understanding the complex effects that context has on QoE needs to be addressed in future research.

13.2 Substantive contributions

I wanted to find out how TV content needs to be presented on mobile devices to be acceptable to a broad audience and how people trade off the most important parameters e.g., size, angular resolution and viewing distance to optimize their visual experience. From the viewpoint of the service provider this is mainly a question about resolution and its appropriate encoding bitrate but my results show that in order to guarantee a satisfying experience the resolution and most importantly the size of the target device have to be considered.

Whereas HDTV aimed at providing a larger picture at the same angular resolution as SDTV my results on mobile TV make a case for providing the same viewing ratio as for SDTV in the living room but at a lower resolution. Many engineers and researchers often scoff at the idea of larger screens without the additional resolution but the benefits of larger screens in terms of angular size trump the reduction in resolution both in the deployment of HD-ready TV sets as well as for mobile TV. Large screens have been around for a while now and can upscale lower resolution content to fill the screen for a more immersive visual experience. Higher resolution content can be made available later to increase the video quality.

13.2.1 Viewing distance

Mobile TV services should be designed for close *viewing distances* between 25cm to 50cm. Distances of 28cm and 32cm were the averages in my studies (1 and 6) in which people rated the acceptability of the overall video quality. I found no adjustment of viewing distance depending on the resolution or the size of the footage. Mobile TV viewing distances seemed to depend more on the posture of people within a given context, e.g. whether they had a backpack with them that was placed on their lap and supported the hand holding the PDA.

13.2.2 Picture size and screen

Many publications and articles mention that mobile TV is small. My research started with focus groups in which a large number of participants initially found the idea of watching TV on a small mobile device questionable due to the assumed small screen size. They did not realise that a mobile device viewed at arm’s length did not require a large size to achieve VR similar to those in their living rooms. They were also worried that a sufficiently large screen might infringe on portability. Clearly, my participants needed to experience video content on a mobile device first hand to realize that size did not impediment an enjoyable experience. At this point I can answer what size is required for mobile. In my experiments size was the most important criterion and the participants preferred to achieve living room TV viewing ratios of approximately 8 for QCIF resolution (see Sec. 13.2.3) content shown at an aspect ratio of 4:3 under lab conditions. At typical viewing distances of around 35cm this can be achieved with a picture height of just over 4cm – possible on many mobile devices. For QoE this represents an important expectation of target users. It might seem trivial that participants expected what they are used to in terms of size but the finding is still surprising as none of the participants made any comment that they expected the same (relative) size on a mobile device or that the demonstrator in the focus groups achieved a comparable size. On the train, size became even more important and viewing ratios larger than 7.8 (at 21ppd) reduced the acceptability. So for mobile contexts a slightly larger minimum screen height of 4.5cm is advisable. Although a viewing ratio of 14H corresponded to the average minimum size at which people found watching QCIF content acceptable in the lab this was only true for about 66% of the audience. Some of my findings on minimal acceptable size that were based on individual clip differences are harder to generalize but revealed a number of additional potential pitfalls such as low contrast and camera movement when targeting minimal sizes. However, considering the large difference between preferred and minimal acceptable size in study 6 (the former was more than 50% larger than the latter) and that extraneous uncontrollable factors might further impair a mobile TV experience as e.g. seen in study 3 designing for minimal acceptable sizes is not advisable in the first place. It would discount the preferences of a substantial part of the audience.

Since people did not adapt their viewing distance I can extrapolate from the average viewing distances of around 30cm in my studies. A target screen heights between 4 and 6cm should make for an acceptable visual experience for most users. On the upper end of the spectrum a 16:9 picture of 10cm height viewed from the largest distance observed in my studies (about 50cm) could provide users with the immersion or *sense of reality* that HDTV was designed to deliver. The required VR of around 5 can be provided by many portable DVD players, laptops and for example the Apple iPad.

Mobile contexts require screens with high luminance and contrast to support use in bright day light. A high screen resolution aids in the depiction of upscaled content and a wide viewing angle will be beneficial for sharing the screen with others. Since screen heights of 4cm can be easily achieved on mobile devices and the resulting viewing ratios are not very different from the TV experience in the living room resolution becomes the major difference. The main change from SDTV to HDTV is an increase in size but the move from SDTV to mobile TV is characterized by a reduction in resolution. The resolution of the display should be well above 115ppi since my results and analysis suggested that a number of participants found low device addressability (115ppi) unacceptable. Apart from small adjustments for small text and small actors my main finding about favourite sizes should hold for video content in general as it depended primarily on resolution.

13.2.3 Resolution

Since people aim at achieving living room viewing ratios the main difference between standard definition and mobile TV is the resolution of the content. Depictions of high spatial frequencies of important detailed objects such as players in XLS and text were affected first by the reduction of resolution and impaired the visual experience. As a rule of thumb, TV content below QCIF (176x144) resolution will not be acceptable to 100% of the audience no matter at what size it is presented or how high it is encoded. If content is not depicted at a sufficiently large picture compared to its angular resolution this may lower acceptability. The most acceptable experience of 4:3 QCIF content on a 200ppi device should be a picture height of 4cm assuming a comparable encoding as in study 1. The angular resolution would be around 20ppd. A general limit for up-scaling video clips regardless of content and shot types was an angular resolution of about 14ppd close to the 11ppd that I derived from Lund's (1993) minimum viewing distances study of large projections of TV content in a dark room.

Content resolution depends to a large degree on the encoding bitrate and in resource constrained settings the question can be framed as to how much more spatial resolution can be gained from increasing the encoding bitrate for a targeted resolution. As I have found in my research the answer to this question can hinge on a number of factors. This is not only due to the different amounts of spatio-temporal information that is present in e.g. sports content *vs.* animation content but is also rooted in the size and rendering of small objects. Depending on the size of the overall picture small players in XLS changed the acceptability of the visual experience as explored in study 5c. In study 2 I showed that moving text was the first part of the picture that was affected due to insufficient encoding bitrates and significantly reduced the acceptability of news content. Future encoders should treat text differently from the rest of the picture and devote more of the encoding budget towards it. Text is a medium in itself and neither content nor shot type specific.

Both size and the available resolution of the content have to be taken into account when making the choice for the best presentation of mobile TV material. My results show that participants' preferences for watching low resolution content depended first on size – they preferred angular sizes between 5.8° (9.8*H*) and 6.6° (8.6*H*) for 120x90 and 168x126 resolution content respectively. These values result in 15ppd and 19ppd angular resolution. From my results it seems plausible that the preferred sizes of content with higher resolution than tested in my experiments will be preferred at even larger angular sizes and angular resolution but the way in which these two parameters are traded-off for higher resolution content requires further research.

13.2.4 Shot types

My results have shown that shot types are important to the understanding of QoE. Apart from XLS, *shot types* were only a concern at the lower limits of acceptable size. MCU and MS could still be presented at smaller sizes than other shot types but their *favourite sizes* did not differ from other shot types. To rely on them in production would only make sense for content that would be shown on displays smaller than 22mm in height (VR>14) – the train results from Study 3, however, indicated that this would be too small for a large part of the audience.

For an acceptable mobile TV experience with sufficient size and resolution as detailed above the XLS was the only shot that justified adaptation through zooming. Usually XLS are used for so-called opening or situating shots instructing viewers where a scene is taking place – the visibility and movement of actors is not important. These "regular" XLS featured in all other content types and had the same favourite sizes as all other shot types. Most complaints about sports XLS targeted a lack of player size and resolution. Study 6 showed that viewers preferred actors at an angular size of 1º but when watching XLS of e.g. ice-hockey they might require still larger sizes because a very small puck is of interest, too.

13.2.5 Zooming

Content adaptation employing zooming only made sense for sports XLS. Content that includes football XLS should be presented at viewing ratios of 11 and larger sizes. For football fans an acceptable experience of watching on mobile devices started at a viewing ratio of 11 with 0.9º angular size players in my studies. Content adaptation employing zooming approaches should target a size of actors in XLS of at least 0.5° as a lower limit but ideally 0.8º at a viewing ratio of 8.5 or smaller with a resolution of 15 pixels in player height. Attaining larger player sizes might further increase the visual experience but when fans had the choice they did not make use of further zooming at the expense of losing contextual information.

Up-scaling the picture on the mobile device can be used to help achieve these sizes – if possible on the device - down to an angular resolution of 17ppd for QCIF content.

The moving of the zooming window introduces more motion, which might adversely affect the viewing experience as described by Holmstrom (2003). Further research is required to identify these optimal pan speeds. Future work would also compare the bespoke moving zoom window tested in my studies with one in which the "safe area" of the broadcast content is cropped off, which is much easier to achieve.

13.3 Methodological contributions

13.3.1 Mobile device viewing ratios - extension to ITU Rec. BT500

The ITU recommendation BT.500 explicitly covers only screen heights equal or greater than 18cm - smaller sizes can be inferred from a graph (*cf.* Figure 58 on page 151). For the covered screen heights it lacks bibliographic references and justification as to why the resulting viewing ratios were chosen. People's preferred viewing distances that change based on screen size might apply to the first guests arriving in cinemas but in the home adjustments of viewing distances in response to screen sizes has not been documented. A recommended viewing distance indicative of people's real viewing conditions in the home (around 3m according to research by Lechner, Jackson and Tanton) would be more appropriate.

Mobile multimedia devices have smaller screens and as long as foldable screens and other projection techniques such as near eye displays are not adopted, this assumed shortcoming might slow down user uptake as the results from my focus groups analysis suggested (Sec. 5.2.). Although once people have experienced mobile TV on screens of 4cm height and larger their worries about an adequate size for the experience that is still portable might be dispelled quickly.

Just as people do not change their living rooms to adjust viewing distances to their TV screen my participants did not alter their posture to change their viewing distance to mobile device screens. According to the ITU video quality measures of SD- and HDTV content do not differ at various viewing ratios but for low resolution mobile content I found a large effect of resolution on the preferred viewing size. The participants' size preferences depended on content resolution – an adjustment that is now possible on a range of devices.

Current ITU recommendations on video quality assessment imply viewing ratios for small screens that are much smaller and result in a poorer overall experience. My results suggested that video quality in mobile services should be evaluated under conditions that resemble people's viewing preferences. Sizes that yielded an angular resolution of 32ppd – identified as optimal picture quality in (Westerink & Roufs 1989) - did not coincide with the participants' favourite sizes but were criticized for being too small. My results on the best viewing conditions were based on lab and field trials and for an assumed fixed viewing distance translated into PVRs. People seek to achieve VRs on mobile devices much larger in terms of the angular size than the ITU recommendations imply.

Objective quality measurements and multimedia models for video content on mobile devices that do not consider these preferred viewing ratios on a mobile device will result in predictions that will not be indicative of people's preferences. Size would be a confounding variable as it was in my first study.

13.3.2 Video quality – one of several factors for QoE

The experience of watching video on small mobile devices cannot be satisfactorily explained through existing video quality models. Video quality research has often focused on human perception and not on user preferences. Barten's SQRI model predicts Westerink & Roufs and Jesty's results on perceived video quality well, which suggested that the preferred viewing ratio would be chosen to attain 32ppd angular resolution. My results showed that this is not the case for low resolution content on mobile devices. The resulting angular size was too small and people preferred to trade-off angular resolution down to 20ppd and below for larger sizes. Objective quality measurements and multimedia models for video content on mobile devices that do not consider the viewing ratio on a target device will make predictions that will not match people's preferences because they discount the significant utility of size and its contribution to the visual experience.

Although research by Yu *et al.* showed that people's video quality judgments are not affected by 3H or 5H presentation of the video these changes do result in different overall experiences and people might prefer one over the other depending on content, shot types, resolution, lighting, contrast, encoding bitrate, desired immersion and social factors. In terms of the visual experience video quality is not a sufficient measure if it does not include a notion of the viewing ratio at which it will be viewed and other preferences that people might have. The over-reliance on video quality as an operationalization for QoE is based in a number of reasons. Video and picture quality especially is well understood. Almost no research has targeted the effects of presentational parameters such aspect ratio and size. Hatada's original research on induced immersion through large displays is missing from many HDTV studies conducted nowadays. It seems as if the long time between the initial ideas for HDTV and its current inceptions broke the chain of scientific referencing. For those that are aware of his work the replication of his elaborate apparatus – based on measuring changes in people's balance – might prove too daunting a task.

13.3.3 Acceptability - a low effort contextualized binary measure

Throughout my experiments I used methods that required little involvement from the participants. For the measure of acceptability the participants had only to monitor whether the visual experience became unacceptable. This was possible while watching the clips without interruptions to the content. None of the participants complained about the method being in the way of following the content. The range of qualities that I used in my experiments did not pose a problem to this binary measure. For higher resolution content acceptability might represent a problem and an alternative label to 'acceptable' should be used. The method delivered stable results and its repeated use allowed for cross-comparison of results from different studies.

High values of acceptability coincided with people's preferences when selecting their preferred trade-off for resolution and size (in study 6). Relying on video quality such as Westerink & Roufs optimal quality of 32ppd to guide (see Sec. 2.6.8) this trade-off would have resulted in sub-optimal results for the overall visual experience. Acceptability, however, framed the rating context specific to a given service and was a good predictor of participants' favourite viewing conditions. Overall this made acceptability a good method to elicit participants' preferences for visual experiences on mobile devices. The instructions contextualized the measure of acceptability as a decision whether to use or not use a service. In future studies I would ask participants to rate the acceptability of the overall visual experience and not mention the term video quality for even clearer instructions as done in study 6.

13.3.4 Qualitative feedback

My findings and the trajectory of my research stress the importance of collecting qualitative feedback. As in previous user-based research e.g. (McCarthy *et al.* 2004a), it helped to explain participants' ratings and disambiguate the effects of different variables on acceptability. Most video quality assessment approaches rely on quantitative or ordinal ratings of one kind that condenses the various contributing dimensions into a single scale. The collection of qualitative feedback in a complaining aloud fashion along with the acceptability ratings allowed for identifying important dimensions with reasonable effort. As seen in study 6 this greatly helped in finding the point at which the confounded variables size and resolution respectively became a concern. This approach was more direct and insightful than debriefs at the end of the experiment, which rely much more on people's memory and make it hard to attribute factors specifically to a certain quality level. This shortcoming became apparent in the analysis of the results of studies 1, 2 and 3.

13.3.5 Assessors

In the definition of QoE the term *expectation* is not sufficiently defined and will need to be filled with more constructive and illustrative meanings in future research. An unspoken expectation of the participants was that mobile TV would be of the same size in relative terms as living room TV.

Expectations about matching qualities in different media dimensions might have caused higher quality audio to reduce the acceptability of the video in study 1. This would also explain why non-native speakers rated news accompanied with text of high visual quality in study 2 less favourably than depictions in which the video quality of the text better matched that of the rest of the picture.

My studies on visual experience showed that people who were interested in football content judged the acceptability of the visual experience differently from people that did not describe themselves as fans. This is line with findings by Jumisko (2005) in which video quality ratings correlated with assessors' degree of interest in the content. The large difference could be due the fact that people who are not interested in football might rate its visual experience more conservatively or punitively. For future research this underscores the value of recruiting participants that are interested in the content or at the very least for including interest in the content as a control variable in the analysis of visual experience.

The people who are most likely to make use of the service involving that content and have much lower thresholds and requirements or a broader more demanding audience? In terms of service adoption it would make most sense to start with people that are interested in the content and to design the service accordingly. As adoption proliferates and performance of the overall system increases the visual quality could be boosted to satisfy the rest of the audience. This would fit well with the typical s-shaped performance curves of service introduction according to Gartner's hype cycle model (Linden & Fenn 2003).

13.3.6 Influence of context

Although my research showed that lab experiments may be a conservative estimate of acceptability of video consumed by people on the move, this was not true for all observed factors. It is important to test preferences in different contexts of use, especially for effects that are not fully understood yet – as in my case image size. Conducting tests of acceptability and user preferences to validate and qualify the results from laboratory results is an essential part of building an understanding of QoE for multimedia services. My results were obtained on trains that induced motion, varying ambient lighting. There are many other conditions that can occur in the field, which might bear different results. My lab trials should be a good approximation for the visual experience of solitary viewing in the home – between 30%-50% of mobile TV field trial participants in (Mason 2006) and (Lloyd *et al.* 2006) used their devices at home as a "personal TV" (Yanqing *et al.* 2007). Social factors, however, might affect the overall user experience in a different manner and should be addressed in future research.

13.3.7 Multi-pronged approach

I have approached my work from different angles while relying on the same material for a range of experimental studies. Besides the standard encoding bitrates and resolution I assessed concerns of viewing ratio, shot types and zooms. I included possible interactions with both audio and text quality. I tested the ecological validity of the results on a train and cross-compared them with results from objective video quality measures. I used qualitative feedback to disambiguate results further and drive subsequent studies both in terms of scrutinizing problems in more detail and develop new foci. Overall this approach allowed for cross-comparisons, disambiguation and the elimination of potentially hidden variables.

13.4 Research update

Most studies about video quality and user experience on mobile devices that were carried out during my work I have reported in the previous chapters. A number of large scale studies have been carried out that trialled mobile TV services– see (Schuurman *et al.* 2009) for an overview. A large expert panel conducted by Schuurman *et al.* confirmed that news, sports and music were the most important content types for mobile TV services. Cartoons came in as the sixth most interesting content type after soap and adult content. Buchinger (2009) carried out a large survey of mobile TV research and identified further work to be carried out on alternatives to headphones, content modification and sharing, channel switching times, improvement of football content, payment models and their interplay with advertising and the usage context.

Bhat *et al.* (2009) proposed a new objective video quality measure MOSp based on temporal and spatial masking information and the mean squared error between the original and compressed video sequences. According to its authors it outperforms another recently proposed method called PSNR+ by Oelbaum *et al.* (2007). It relies on edge detection in local regions of the image and its predictions are highly correlated with subjective MOS results. This matches well with my findings on text. The rendering of text in videos contained many pronounced edges and the legibility of text was very important for acceptability in my studies.

13.5 Future trends

Whether or not the mobile consumption of multimedia content will be popular and mobile TV see widespread adoption is of not much concern to the work presented here. Optimising video content for delivery and consumption with resource constraints will be of interest for the design of many services -

downloading content, secondary displays for home entertainment, peer-to-peer TV and delivering content to all sorts of display devices. Displays will continue to increase in addressability, luminance, contrast and response times all of which will determine how well spatio-temporal information can be mediated. The resolution will reach human perceptual threshold at which the pixels cannot be resolved at regular operating distances providing a 'perfect canvas' even at the shortest possible viewing distances. Improved video encoders will further reduce the amount of encoding bitrates required to render a given resolution and will result in higher perceived quality. The question whether content will be available at the same resolution to be displayed natively is not so much a technical but an economic question. Considering the large gap between what people find acceptable in terms of spatial resolution and their discriminatory threshold it is not clear in how far the added value of higher resolution will be met by people's willingness to pay or to accept a more limited offering in terms of content, e.g. channels. Most likely the native resolution of displays will exceed that of the content distributed for mobile devices. Larger and higher resolution display will make the question of how far to upscale video content more important. The most recent arrival - Apple's iPad - is another harbinger of assumed consumption of video content on mobile devices in the home. Another trend that relates to upscaling is the proliferation of low resolution video recordings from mobile devices such as phones. This content will be presented in an upscaled version on regular TVs. Memory will continue to drop in price and make full PVR functionality with ample amounts of storage capacity available on mobile TV devices. Cropping algorithms that enlarge parts of the content might become a solution if the content depicted on mobile TV screens is too small for the viewer. Mobile phones with video camera capabilities might make for a very different mobile TV experience if peers or groups of people start providing each other with video clips on the go.

13.6 Directions for future work

Understanding the experiences afforded by technology in the field and the value that people attribute to these will be a challenge. Future research needs to devise models for QoE that go beyond audio and video quality as they exist today. Visual experience could prove a valuable component in this definition of QoE that should supersede video quality. Models of visual experience would need to include viewing and aspect ratios as independent variables as well as the contribution of e.g. 2D or 3D rendering, sound and other experiential parameters still to be discovered. Higher levels of visual experience than acceptable need to be evaluated and other concepts that capture experience from different angles such as immersion, annoyance and enjoyment need to be addressed in QoE research. Some factors, which might have an effect on the experience of multimedia consumption on mobile devices that I did not address in my research, are fatigue (due to short viewing distances, and possibly smaller angular sizes), contrast, artefacts due to data loss, errors and loss of synchronization.

Practitioners will need to investigate the best trade-off in terms off at which nominal resolution to encode content especially when up-scaling content on high resolution displays might occur. A method to derive the detail in terms of the visual frequencies in a picture and to match that with appropriate encoding bitrates for specific local regions such as text would be useful. It would be interesting to cross-check my acceptability scores of news with MOSp values of the same content. My work cannot answer what encoding bitrate is required for a given resolution for a given codec. This is complicated by the different spatio-temporal information contained in the video. The main problem challenge lies in finding the best point at which an increase in encoding bitrate only marginally increases the achieved experience.

References

Odyssey software inc. CFCOM (2003). http://www.odysseysoftware.com/

Agarwal, G., Anbu, A., & Sinha, A. (2003). A Fast Algorithm To find The Region-Of-Interest in the Compressed MPEG Domain. In *Proceedings International Conference on Multimedia and Expo* (pp. 133-136).

Agrafiotis, D., Canagarajah, N., & Bull, D. R. (2003). Perceptually Optimised Sign Language Video Coding. In *Proceedings of the 2003 10th IEEE International Conference on Electronics, Circuits and Systems* (pp. 623-626).

Agro, L. (2005). Microdisplay Emotions. http://www.leeander.com/2005/04/microdisplay_emotions.html

Ahmed, K., Kooij, R., & Brunnström, K. (2006). Perceived quality of channel zapping. In *ITU-T Workshop on QoE/QoS*.

Ajzen, I. & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ, USA: Prentice-Hall.

Aldrich, S. E., Marks, R. T., Lewis, J. M., & Seybold, P. B. (2000). *What kind of the Total Customer Experience Does Your E-Business Deliver?* Patricia Seybold Group.

Aldridge, R. P., Hands, D. S., Pearson, D. E., & Lodge, N. K. (1995). Continuous assessment of digitally-coded television pictures. *IEE Proceedings - Vision, Image and Signal Processing, 145,* 116-123.

American National Standards Institute (ANSI) (1988). *American national standard for human factors engineering of visual display terminal workstations* (Rep. No. ANSI/HFS Standard No.100-1988). Santa Monica, CA: The Human Factors Society Inc.

Ankrum, D. R. (1996). Viewing Distance at Computer Workstations. *Work Place Ergonomics,* 10-12.

Apple (2005). QuickTime [Computer software].

Apteker, R. T., Fisher, A. A., Kisimov, V. S., & Neishlos, H. (1994). Distributed multimedia: user perception and dynamic QoS. In *Proceedings of SPIE* (pp. 226-234).

Ardito, M. (1994). Studies on the Influence of Display Size and Picture Brightness on the Preferred Viewing Distance for HDTV programs. *SMPTE, 103,* 517-522.

Ardito, M., Gunetti, M., & Visca, M. (1996). Influence of display parameters on perceived HDTV quality. *IEEE Transactions on Consumer Electronics, 42,* 145-155.

Assfalg, J., Bertini, M., Colombo, C., & Del Bimbo, A. (2003). Automatic extraction and annotation of soccer video highlights. In *Image Processing, 2003.Proceedings*.

Attneave, F. (1962). Perception and related areas. In S.Koch (Ed.), *Psychology: A study of a science* (pp. 619-659). New York, NY: McGraw-Hill.

Avisynth (2005). http://www.avisynth.org/

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology, 3,* 87-103.

Bailey, I. L. & Lovie, J. E. (1976). New Design Principles for Visual Acuity Letter Charts. *American Journal of Optometry & Physiological Optics, 53,* 740-745.

Baker, K. (2006). A Dodo of History: Interactive TV. In.

Baldwin, M. W., Jr. (1940). The Subjective Sharpness of Simulated Television Images. *Proceedings of the IRE, 28,* 458-468.

Barber, P. J. & Laws, J. V. (1994). Image Quality and Video Communication. In R. Damper, W. Hall, & J. Richards (Eds.), *Proc of IEEE Int'l Symposium on Multimedia Technologies & their Future Applications* (pp. 163-178). London, UK: Pentech Press.

Barten, P. G. J. (1990). Evaluation of subjective image quality with the square-root integral method. *Journal of Optical Society of America, 7,* 2024-2031.

Bathia, S., Lakshminarayanan, V., Samal, A., & Welland, G. V. (1995). Human face perception in degraded images. *Journal of Visual Communication and Image Representation, 6,* 280-295.

Bauer, B. & Patrick, A. S. (2002). A human factors extension to the seven-layer osi reference model. http://www.iit.nrc.ca/~patricka/OSI/10layer.pdf

Beauregard, R., Younkin, A., Corriveau, P., Doherty, R., & Salskov, E. (2007). Assessing the Quality of User Experience. *Intel Technology Journal: Designing Technology with People in Mind, 11*.

Beerends, J. G. & de Caluwe, F. E. (1999). The influence of video quality on perceived audio quality and vice versa. *Journal Audio Eng.Soc., 47,* 355-362.

Benbasat, I. & Barki, H. (2007). Quo Vadis, TAM? *Association for Information Systems, Journal of, 8,* 211-218.

Bennett, A. G. (1965). Ophthalmic test types. *Br.J.Physiol.Opt., 22,* 238-271.

Bergfeld Mills, C. & Weldon, L. J. (1987). Reading Text from Computer Screens. *ACM Computing Surveys, 19,* 329-358.

Bertram, R. & Steinmetz, R. (1997). Scalability of audio quality for networked multimedia environments. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems '97* (pp. 294-301). Ottawa, Ont Canada: IEEE Comput. Soc.

Bhat, A., Richardson, I., & Kannangara, S. (2009). A novel perceptual quality metric for video compression. In *Proceedings of the 27th conference on Picture Coding Symposium* (pp. 141-144).

Birkmaier, C. (2000). Understanding Digital: Advanced Theory. In M.Silbergleid & M. Pescatore (Eds.), *The Guide to Digital Television* (United Entertainment Media.

Blythe, M. A., Overbeeke, K., Monk, A. F., & Wright, P. C. (2003). *Funology*. (vols. 3) Springer.

Borg, G. (1982). A category scale with ratio properties for intermodal and inter-individual comparisons. In H.G.Geissler & P. Petzold (Eds.), *Psychophysical judgment and the process of perception* (pp. 25-34). Berlin, Germany: Deutscher Verlag der Wissenschaften.

Bouch, A. & Sasse, M. A. (1999). Network QoS: What do users need? In *Proceedings of IDC'99*.

Bouch, A. & Sasse, M. A. (2001). Why Value is Everything: A user-centred approach to Internet Quality of Service and Pricing. In L. Wolf, D. Hutchison, & R. Steinmetz (Eds.), *Proceedings of 9th International conference on Quality of Service (IWQoS'01)* (pp. 49-72). Springer.

Bouch, A., Sasse, M. A., & de Meer, H. (2000). Of Packets and People: A User-Centred Approach to Quality of Service. In *Proceedings of IWQoS 2000* (pp. 189-197).

Bouch, A. (2001). *A User-centered Approach to Network Quality of Service and Charging.* PhD Thesis Unpublished thesis: University College London.

Breakwell, G. M. (2000). Interviewing. In G.M.Breakwell, S. Hammond, & C. Fife-Schaw (Eds.), *Research Methods in Psychology* (pp. 239-250). London: SAGE.

Briel, R. (2008). MFD hands back German T-DMB licence. http://www.broadbandtvnews.com/?p=4682

Buchinger, S. (2009). A comprehensive view on user studies: survey and open issues for mobile TV. In *Proceedings of the seventh european conference on European interactive television conference* (pp. 179-188).

Buchinger, S. & Hlavacs, H. (2008). A Low Start for Mobile Video Patching. In *EuroFGI IA.7.6 Workshop on Socio-Economic Aspects of Future Generation Internet 2008*.

Campbell, A., Coulson, G., Hutchinson, D., & Leopold, H. (1993). Integrated Quality of Service for Multimedia Communications. In *Proc.IEEE INFOCOM'93* (pp. 732-739). IEEE.

CCITT (1988). *Quality of Service and dependability vocabulary* (Rep. No. E.800).

Cesar, P., Bulterman, D. C. A., Geerts, D., Jansen, J., Knoche, H., & Seager, W. (2008). Enhancing Social Sharing of Videos: Fragment, Annotate, Enrich, and Share. In *Proceedings of ACM Multimedia 2008*.

Chapanis, A. & Scarpa, L. C. (1967). Readability of Dials at Different Distances with Constant Visual Angle. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 9,* 419-425.

Chorianopoulos, K. (2004). *Virtual Television Channels: Conceptual Model, User Interface Design and Affective Usability Evaluation.* PhD Unpublished thesis: Athens University of Economics and Business.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *PNAS, 102,* 12629-12633.

Chuang, S. L. & Haines, R. F. (1993). A Study of Video Frame Rate on the Perception of Compressed Dynamic Imagery. In *Society for Information Display, 1993 International Symposium*.

Collins, C., O'Meara, D., & Scott, A. B. (1975). Muscle strain during unrestrained human eye movements. *Journal of Physiology, 245,* 351-369.

Cooper, A., Reimann, R., & Cronin, D. (2007). *About Face 3 - The Essentials of Interaction Design*. Indianapolis, IN, USA: Wiley Publishing, Inc.

Corey, G. P., Clayton, M. J., & Cupery, K. N. (1983). Scene Dependence of image quality. *Society of Photographic Scientists and Engineers, 27,* 9-13.

Cowan, M. (2002). *Digital Cinema Resolution - Current Situation and Future requirements* http://www.etconsult.com/papers/Technical%20Issues%20in%20Cinema%20Resolution.pdf: Entertainment Technology Consultants.

Crano, W. D. (1977). Primacy Versus Recency in Retention of Information and Opinion Change. *The Journal of Social Psychology, 101,* 87-96.

Daintith, J. (2004). Resolution: A Dictionary of Computing. Encyclopedia.com

Dal Lago, G. (2006). Microdisplay Emotions. http://www.srlabs.it/articoli_uk/ics.htm

Dan, A., Sitaram, D., & Shahabuddin, P. (1994). Scheduling policies for an on-demand video server with batching. In *Proc.of ACM Multimedia* (pp. 15-23).

Dancyger, K. (2008). *The Technique of Film and Video Editing: History, Theory, and Practice*. Focal Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13,* 319-340.

de Koning, T. C. M., Veldhoven, P., Knoche, H., & Kooij, R. (2007). Of MOS and men: bridging the gap between objective and subjective quality measurements in mobile TV. In *Proceedings of Multimedia on Mobile Devices 2007*.

de Ridder, H., Blommaert, F. J., & Fedorovskaya, E. A. (1995). Naturalness and image quality: Chroma and hue variation in color images of natural scenes. In *Proc.SPIE* (pp. 51-61).

de Ridder, H. & Hamberg, R. (1997). Continuous assessment of image quality. *SMPTE Journal,* 123-128.

de Vries, E. (2006). Renowned Philips picture enhancement techniques will enable mobile devices to display high-quality TV images.
http://www.research.philips.com/technologies/display/picenhance/index.html

Delbecq, A. L. & Van de Ven, A. H. (1971). A Group Process Model for Problem Identification and Program Planning. *Journal of Applied Behavioral Science, VII,* 466-491.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group Techniques for Program Planners*. Glenview, IL: Scott Foresman and Company.

Deloitte (2006). *Deloitte: 3G, Mobile TV Disappoint - MVNOs Fly* (Rep. No. http://www.accessmylibrary.com/article-1G1-141865700/deloitte-3g-mobile-tv.html).

Dey, A. K., Abowd, G. D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *International Journal on Human-Computer Interaction, 16,* 97-166.

Diamant, L. (1989). *The Broadcast Communications Dictionary*. (3rd ed.) Greenwood Press.

Dorr, A. (1980). When I was a child I thought as a child. In S.B.Whithey & R. P. Abeles (Eds.), *Television and social behaviour: Beyond violence and children* (Lawrence Earlbaum Associates.

Drucker, P., Glatzer, A., De Mar, S., & Wong, C. (2004). SmartSkip: Consumer level browsing and skipping of digital video content. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves* (pp. 219-226). New York, NY, USA: ACM Press.

Drucker, P. (1999). *Management*. Butterworth-Heinemann.

Duncanson, J. P. & Williams, A. D. (1973). Video conferencing: Reactions of Users. *Human Factors, 15,* 471-485.

EBU (2004). *EBU Technical Recommendation R112 - 2004* (Rep. No. EBU Technical Recommendation R112 - 2004). Geneva, Switzerland: EBU.

EBU (2008). *Safe areas for 16:9 television production* (Rep. No. Technical Recommendation R95-2008). Geneva, Switzerland: EBU.

Edgeworth, F. Y. (1881). *Mathematical Psychics: An Essay on the Application of Mathematics to Moral Sciences*. (reprinted ed.) M. Kelly (1954).

Ehn, P. & Kyng, M. (1991). Cardboard Computers: Mocking-it-up or Hands-on the Future. In J.Greenbaum & M. Kyng (Eds.), *Design at Work: Cooperative Design of Computer Systems* (Hillsdale, NJ: Lawrence Erlbaum.

Empririx. (2004). Assuring QoE on Next generation Networks.

Engeldrum, P. G. (2001). Psychometric Scaling: Avoiding the Pitfalls and Hazards. In *IS&T's 2001 PICS Conference Proceedings* (pp. 101-107).

Engle, F. L. (1971). Visual Conspicuity, Directed Attention and Retinal Locus. *Vision Research, 11,* 563.

Ericsson (2006). *Video streaming quality measurement with VSQI* (Rep. No. Technical Paper). Ericsson.

ETSI (2005). Digital Video Broadcasting (DVB); DVB-H Implementation Guidelines. http://webapp.etsi.org/action/PU/20050301/tr_102377v010101p.pdf

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Feistinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fink, D. G. (1955). Color television vs. color motion pictures. *SMPTE, 64,* 281-290.

Fisher, R. F. (1977). The force of contraction of the human ciliary muscle during accommodation. *Journal of Physiology, 270,* 51-74.

Flavell, J. H., Flavell, E. R., Green, F. L., & Korfmacher, J. E. (1990). Do young children think of television images as pictures or real objects? *Broadcasting and Electronic Media, Journal of, 34,* 399-419.

Fredrickson, B. & Kahneman, D. (1993). Duration Neglect in Retrospective Evaluations of Affective Episodes. *Journal of Personality and Social Psychology, 65,* 45-55.

Frieser, H. & Biedermann, K. (1963). Experiments on image quality in relation to the modulation transfer function and graininess of photographs. *Photographic Science and Engineering, 7*.

Frohlich, D. M., Thomas, P., Hawley, M., & Hirade, K. (1997). Future personal computing. *Personal Technologies, 1,* 1-5.

Fujine, T., Yoshida, Y., & gino, M. (2008). The relationship between preferred luminance and TV screen size. In S. P. Farnand & F. Gaykema (Eds.), *Proceedings of the SPIE* (pp. 68080Z-68080Z-12).

Fujio, T. (1985). High-Definition Television Systems. *Proceedings of the IEEE, 73,* 646-655.

Fulton Suri, J. (2009). Designing Experience: Whether to Measure Pleasure or Just Tune In? In W.S.Green & P. W. Jordan (Eds.), *Pleasure With Products: Beyond Usability* (London, UK: Taylor & Francis.

Gauntlett, D. & Hill, A. (1999). *TV Living: Television, Culture and Everyday Life.* Routledge.

Ghinea, G. & Chen, S. Y. (2004). The impact of cognitive styles on perceptual distributed multimedia quality. In.

Ghinea, G. & Thomas, J. P. (1998). QoS Impact on User Perception and Understanding of Multimedia Video Clips. In *Proceedings of ACM Multimedia '98*.

Ghinea, G. & Thomas, J. P. (2001). Crossing the Man-Machine Divide: A Mapping Based on Empirical Results. *Journal of VLSI Signal Processing, 29,* 139-147.

Gibbs, A. (2004). Focus Groups. *Social Research Update*.

Glaser, B. G. & Strauss, A. L. (1999). *The discovery of grounded theory: Strategies for qualitative research*. New Brunswick, USA: Aldine Transaction.

Goodman, D. & Nash, R. (1982). Subjective Quality of the Same Speech Transmission Conditions in Seven Different Countries. *IEEE Transactions on Communications, 30,* 642-654.

Goulev, P. (2004). Affective Ware. http://www.iis.ee.ic.ac.uk/~p.goulev/download/affectiveware.zip

Gouriet, G. G. (1958). Discussion Before th Radio and Telecommunication Section, 19th February 1958. *Proc.Institution of Electrical Engineering, 105B,* 435.

Greenbaum, J. & Kyng, M. (1991). *Design at Work - Cooperative design of Computer Systems*. Lawrence Earlbaum.

Grinter, R. E. & Eldridge, M. (2001). y do tngrs luv 2 txt msg? In *Proceedings of the seventh conference on European Conference on Computer Supported Cooperative Work* Kluwer Academics Publisher.

Grudin, J. (2005). Three Faces of Human-Computer Interaction. *Annals of the History of Computing, IEEE, 27,* 46-62.

Guardian (2005). Romantic drama in China soap opera only for mobile phones. Guardian Newspapers Limited Available: http://www.buzzle.com/editorials/6-28-2005-72274.asp

Gulliksen, J., Blomkvist, S., & Straub, D. W. (2003). Engineering the HCI profession or softening development processes. In J. Jacko & C. Stephanidis (Eds.), *Human-Computer Interaction.Theory and Practice (Part I).Volume 1 of the Proceedings of HCI International 2003* (pp. 118-122).

Gwinn, E. & Hughlett, M. (2005). Mobile TV for your cell phone. Chicago Tribune Available: http://home.hamptonroads.com/stories/story.cfm?story=93423&ran=38197

Hall, E. T. (1966). *The hidden dimension*. New York, NY, USA: Doubleday & Company Inc.

Hands, D. S. (2004). A Basic Multimedia Quality Model. *IEEE Transactions on Multimedia, 6,* 806-816.

Hands, D. S. & Avons, S. E. (2001). Recency and duration neglect in subjective assessment of television picture quality. *Applied Cognitive Psychology, 15,* 639-657.

Hands, D. S., Jacobs, R., & Chang, K. (2007). Price-dependent quality: examining the effects of price on multimedia quality requirements. In *Proceedings of Human Vision and Electronic Imaging XII.*

Hansen, V. (2005). *Designing for interactive television: version 1.0.* British Broadcasting Corporation (BBC).

Hardman, V., Sasse, M. A., Handley, M., & Watson, A. (1995). Reliable Audio for Use over the Internet. In *Proceedings of INET'95* (pp. 171-178). Reston, VA: ISOC.

Harper, R., Regan, T., & Rouncefield, M. (2008). Taking Hold of TV: Learning From the Literature. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments* (pp. 79-86).

Hart, W. M. (1987). The temporal responsiveness of vision. In R.A.Moses, W. Hart, & W. M. Hart (Eds.), *Adler's Physiology of the eye, Clinical Application* (St. Louis, MO, USA: The C. V. Mosby Company.

Haskell, B. G., Puri, A., & Netravali, A. N. (1997). *Digital video: an introduction to MPEG-2*. Kluwer Academic Publishers.

Hatada, T., Sakata, H., & Kusaka, H. (1980). Psychophysical analysis of the 'sensation of reality' induced by a visual wide-field display. *SMPTE, 89,* 560-569.

Hauske, G., Stockhammer, T., & Hofmaier, R. (2003). Subjective Image Quality of Low-Rate and Low-Resolution Video Sequences. In *Proceedings of the 8th International Workshop on Mobile Multimedia Communications*.

Hearnshaw, D. (1999). *Desktop Videoconferencing for Tutorial Support.* Unpublished thesis: University College London.

Hellström, G. (1997). Quality measurement on Video Communication for Sign Language. In *Proc.Of 16th International Symposium on Human Factor inTelecommunications* (pp. 217-224).

Hewett, T. T., Baecker, R., Card, S. K., arey, R., asen, J. G., Mantei, M. M. et al. (1992). ACM SIGCHI Curricula for Human-Computer Interaction. http://sigchi.org/cdg/cdgR.html

Hogarth, R. M. (1980). *Judgement and Choice*. Wiley.

Holmstrom, D. (2003). *Content based pre-encoding video filter for mobile TV.* Unpublished thesis: Umea University, http://exjob.interaktion.nu/files/id_examensarbete_5.pdf.

Hsu, C. & Hefeeda, M. (2009). Bounding Switching Delay in Mobile TV Broadcast Networks. In *Proc.of ACM/SPIE Multimedia Computing and Networking Conference*.

Hughes, J., O'Brien, J., & Rodden, T. (1998). Understanding Technology in Domestic Environments: Lessons for Cooperative Buildings. In *Proceedings of the First International Workshop on Cooperative Buildings (CoBuild'98)* (pp. 248-261). Heidelberg, Germany: Springer.

Husemann, D. (2001). Pervasive computing: Hogwarts, StarTrek, reality and back. *Computer Networks, 35,* 373-375.

IBM (2002). Functions of mobile multimedia QOS control. http://www.trl.ibm.com/projects/mmqos/system_e.htm

International Standardization Organization (ISO) (1999). *Human-centred design processes for interactive systems.* (Rep. No. ISO 13407:1999). Geneva, Switzerland: International Standardization Organization.

ITU-R (2004). *Methodology for the subjective assessment of the quality of television pictures.* (Rep. No. BT.500-11).

ITU-T (1996). *Information technology - Open Distributed Processing - Reference Model: Foundation* (Rep. No. Recommendation X.902).

ITU-T (1999). *Subjective audiovisual quality assessment methods for multimedia applications* (Rep. No. ITU-T Recommendation P.911).

ITU-T (2001). *End-user multimedia QoS categories* (Rep. No. G.1010).

ITU-T (2004). *P.800.Methods for subjective determination of transmission quality* (Rep. No. ITU-T P.800).

ITU-T (2008). *Amendment 2: New definitions for inclusion in Recommendation P.10/G.100* (Rep. No. Rec. P.10/G.100).

Jacob, R. J. K. & Karn, K. S. (2004). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises (Section Commentary). In J.Hyona, R. Radach, & H. Deubel (Eds.), *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movement* (Oxford: Elsevier Science.

Jesty, L. C. (1958). The relation between picture size, viewing distance and picture quality. *Proc.Institution of Electrical Engineering, Part B 105,* 425-439.

Johnson, J. (1958). Image Intensifier Symposium. In *AD 220160*.

Joost (2009). www.joost.com

Jordan, P. W. (2000). *Designing Pleasurable Products: An Introduction to Human Factors*. London, UK: Taylor & Francis.

Jumisko, S. (2005). Effect of TV Content in Subjective Assessment of Video Quality on Mobile Devices. In R. Creutzburg & J. H. Takala (Eds.), *Proceedings of SPIE, volume 5684.Multimedia on Mobile Devices* (pp. 243-254).

Jumisko-Pyykkö, S. & Häkkinen, J. (2006). "I would like see the face and at least hear the voice": Effects of Screen Size and Audio-video Bitrate Ratio on Perception of Quality in Mobile Television. In *Proceedings of EuroITV '06*.

Jumisko-Pyykkö, S., Häkkinen, J., & Nyman, G. (2007). Experienced Quality Factors - Qualitative Evaluation Approach to Audiovisual Quality. In *Proceedings of IST/SPIE conference Electronic Imaging, Multimedia on Mobile Devices 2007*.

Jumisko-Pyykkö, S. & Hannuksela, M. (2008). Does Context Matter in Quality Evaluation of Mobile Television. In *Proc.of mobile HCI*.

Jumisko-Pyykkö, S. & Väänänen-Vainio-Mattila, K. (2006). The Role of Audiovisual Quality in Mobile Television. In *Proceedings of Second International Workshop in Video Processing and Quality Metrics for Consumer Electronics*.

Kaasinen, E. (2005). *User acceptance of mobile services – value, ease of use, trust and ease of adoption.* Unpublished thesis: VTT publications 566, Helsinki, Finland.

Kahneman, D. (2003). Experience Utility and Objective Happiness. In D.Kahneman & A. Tversky (Eds.), *Choices, Values and Frames* (New York, NY, USA: Russell Sage Foundation.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of Experienced Utility. *Quarterly Journal of Economics, 112,* 375-405.

Kato, S., Boon, C. S., Fujibayashi, A., Hangai, S., & Hamamoto, T. (2005). Perceptual Quality of Motion of Video Sequences on Mobile Terminals. In *Proceedings of the IASTED International Conference* (pp. 442-447).

Kaufman, J. E. & Christensen, J. F. (1987). *IES Lighting Handbook: 1987 Application Volume*. New York, NY, USA: Illuminating Engineering Society of North America.

Kell, R. D., Bedford, A. V., Fredendall, G. L., & Frederikson, L. (1940). Determination of Optimum Number of Lines in Television Systems. *RCA Review, 5,* 8-30.

Kell, R. D., Bedford, A. V., & Trainer, M. A. (1934). Experimental Television System - The Transmitter. *Proceedings of the IRE, 22,* 1265.

Kelly, D. H. (1979). Motion and vision. II. Stabilized spatio-temporal threshold surface. *J.Opt.Soc.Am., 69*.

Kelly, S. (2006). Content challenge for mobile TV. BBC Home Available: http://news.bbc.co.uk/1/hi/programmes/click_online/4724068.stm

Kies, J. K., Williges, R. C., & Rosson, M. B. (1996). *Controlled Laboratory Experimentation and Field Study Evaluation of Video Conference for Distance Learning Applications* (Rep. No. HCIL 96-02). Virginia Tech.

Kingslake, R. (1963). *Lenses on Photography*. New York, NY: Barnes & Co.

Knoche, H., de Meer, H., & Kirsh, D. (2005). Compensating for Low Frame Rates. In *CHI '05 extended abstracts on Human factors in computing systems* (pp. 1553-1556).

Knoche, H. & McCarthy, J. (2004). Mobile Users' Needs and Expectations of Future Multimedia Services. In *Proceedings of the WWRF12*.

Knoche, H. & McCarthy, J. (2005). Design Requirements for Mobile TV. In *Proceedings of Mobile HCI* (pp. 69-76).

Knoche, H. & Sasse, M. A. (2004). *Blueprint for Focus Groups* (Rep. No. TN01-1.1-02_UCL_MAESTRO_V3.0).

Kopf, S., Lampi, F., King, T., & Effelsberg, W. (2006). Automatic Scaling and Cropping of Videos for Devices with Limited Screen Resolution. In *Proceedings of the 14th annual ACM international conference on Multimedia* (pp. 957-958).

Kortum, P. & Sullivan, M. (2004). Content is King: The Effect of Content on the Perception of Video Quality. In *Human Factors and Ergonomics Society Annual Meeting Proceedings, Perception and Performance* (pp. 1910-1914).

Koskinen, I. & Repo, P. (2006). *Personal Technology in Public Places - Face and Mobile Video* (Rep. No. 94 - 2006). Helsinki, Finland: National Consumer Research Centre.

KPMG (2006). *Consumers and Convergence*
Challenges and opportunities in meeting next generation customer needs.

Kubota, S., Shimada, A., Okada, S., Nakamura, Y., & Kido, E. (2006). Television Viewing Conditions at Home. *Journal of the Institute of Image Information and Television Engineers, 60,* 597-603.

Kumar, A. (2007). *Mobile TV: DVB-H, DMB, 3G Systems and Rich Media Applications*. Burlington, MA, USA: Focal Press.

Lienhart, R. (1999). Comparison of Automatic Shot Boundary Detection Algorithms. In *Proc.of SPIE Storage and Retrieval for Image and Video Databases VII* (pp. 290-301).

Linden, A. & Fenn, J. (2003). *Understanding hype cycles* (Rep. No. Strategic Analysis Report R-20-1971).

Lloyd, E., Maclean, R., & Stirling, A. (2006). *Mobile TV - results from the BT Movio DAB-IP pilot in London* (Rep. No. Technical Review 306). EBU.

Logan, R. J. (1994). Behavioral and emotional usability: Thomson consumer electronics. In M.E.Wiklund (Ed.), *Usability in Practice* (pp. 59-82). New York, NY, USA.

Lombard, M., Grabe, M. E., Reich, R. D., Campanella, C., & Ditton, T. B. (1996). Screen Size and viewer responses to television: A review of research. In *Annual Conf.of the Assoc.for Education in Journalism and Mass Communication*.

Lombard, M., Reich, R. D., Grabe, M. E., Bracken, C. C., & Ditton, T. B. (2000). Presence and Television: The role of screen size. *Human Communication Research, 26,* 75-98.

Lund, A. M. (1993). The Influence of Video Image Size and Resolution on Viewing-Distance Preference. *SMPTE, 102,* 406-415.

Lunt, P. & Livingstone, S. (1996). Rethinking the Focus Group in Media and Communications Research. *Journal of Communication, 46,* 79-98.

Luther, A. C. (1996). *Principles of Digital Audio and Video*. Boston, London: Artech House Publishers.

Mäki, J. (2005). *Finnish Mobile TV pilot* Research International Finland.

Marks, L. E. & Algom, D. (1998). Psychophysical Scaling. In M.H.Birnbaum (Ed.), *Measurement, Judgment, and Decision Making* (Second ed., Academic Press.

Martin, E. R. (1985). HDTV - A DBS Perspective. *Selected Areas in Communications, IEEE Journal on, 3,* 76-86.

Mason, S. (2006). *Mobile TV - Results from the BT Movio DAB-IP trial in Oxford* EBU Technical Review.

Masoodian, M., Apperley, M., & Frederikson, L. (1995). Video support for shared workspace interaction: an empirical study. *Interacting with Computers, 7,* 237-253.

Masry, M. A. & Hemami, S. S. (2004). CVQE: A Metric for Continuous Video Quality Evaluation at Low Bit Rates. In.

Mauss, I. B., Wilhelm, F. H., & Gross, J. J. (2004). Is there less to social anxiety than meets the eye? *Cognition and Emotion, 18,* 631-662.

McCarthy, J. (2003). *The Analysis and Interpretation of Eye Motion* (Rep. No. Research Report University College London).

McCarthy, J., Sasse, M. A., & Miras, D. (2004a). Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video. In *Proc.CHI* (pp. 535-542).

McCarthy, J. & Wright, P. (2004). *Technology as experience*. Cambridge, Massachusetts: MIT Press.

McCarthy, J., Miras, D., & Knoche, H. (2004b). *TN01-1.1.03_UCL_MAESTRO_bandwidth_study_V02*.

McCloud, S. (1994). *Understanding Comics*. Harpers.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 254,* 746-748.

McLuhan, M. (1964). *Understanding Media, The Extensions of Man*. Cambridge, MA, USA: MIT Press.

McVey, G. F. (1970). Television: Some viewer-display considerations. *Educational Technology Research and Development, 18,* 277-290.

Mellers, B. J. & Cook, A. D. J. (1996). The role of task and context in preference measurement. *Psychological Science, 7.*

Mertz, P. & Gray, F. (1934). Theory of Scanning and Its Relation to Characteristics of Transmitted Signal in Telephotography and Television. *Bell System Technical Journal, 13,* 464-468.

Meyer, D. (2007). BT ditches mobile TV service. http://news.zdnet.co.uk/communications/0,1000000085,39288247,00.htm?r=1

Miller, R. B. (1968). Response time in man-computer conversational transactions. In *Proceedings of the Fall Joint Computer Conference* (pp. 267-277).

Mitsuhashi, T. (1982). *Scanning specifications and picture quality* (Rep. No. NHK Techn. Monograph). Tokyo: Japan Broadcasting Corporation.

Mitsuhashi, T. & Yuyama, I. (1991). Recent Advances in HDTV - Present State and Future Development of HDTV. In *Image Management and Communication (IMAC) in Patient Care: New Technologies for Better Patient Care* (pp. 212-217).

Möller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications*. Boston, MA, USA: Kluwer Academic Publishers.

Muir, L. & Richardson, I. E. G. (2002). Video telephony for the deaf: Analysis and development of an optimised video compression product. In *Proceedings of the tenth ACM international conference on Multimedia* (pp. 650-652).

Mullin, J., Smallwood, L., Watson, A., & Sasse, M. A. (2001). New techniques for assessing audio and video quality in real-time interactive communications. In.

Musgrave, G. (2001). Legibility of Projected Information.
www.conceptron.com/articles/pdf/legibility_of_projected_information.pdf

Nathan, J. G., Anderson, D. R., Field, D. E., & Collins, P. (1985). Television viewing at home: Distances and visual angles of children and adults. *Human Factors, 27,* 467-476.

Nemethova, O., Ries, M., Dantcheva, S., Fikar, S., & Rupp, M. (2005). Test Equipment of Time-Variant Subjective Perceptual Video Quality in Mobile Terminals. In *Proc.of HCI*.

Nemethova, O., Zahumensky, M., & Rupp, M. (2004). Preprocessing of Ball Game Video-Sequences for Robust Transmission over Mobile Networks. In *Proceedings of the CIC 2004 The 9th CDMA International Conference*.

Neuman, W. R. (1988). The Mass Audience looks at HDTV: An early experiment. In *Research Panel National Association of Broadcasters Annual Convention*.

Neumann, W. R., Crigler, A. N., & Bove, V. M. (1991). Television Sound and Viewer Perceptions. In *Proc.Joint IEEE/Audio Eng.Soc.Meetings* (pp. 101-104).

Nickerson, R. S. & Landauer, T. K. (1997). Human Computer Interaction: Background and Issues. In M.Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., pp. 3-32). Amsterdam, Holland: Elsevier.

Nokia (2004). *Quality of Experience (QoE) of mobile services: Can it be measured and improved?* (Rep. No. http://www.nokia.com/BaseProject/Sites/NOKIA_MAIN_18022/CDA/Categories/AboutNokia/Press/WhitePapers/Networks/_Content/_Static_Files/whitepaper_qoe_net.pdf).

Nokia (2006). *Abertis Telecom, Nokia and Telefonica Moviles unveil results of first digital mobile TV pilot in Spain* (Rep. No. http://www.mobiletv.nokia.com/news/showPressReleases/?id=73).

Norman, D. (2005). Human-Centered Design Considered Harmful. *Interactions, 12,* 14-19.

Norman, D. A. (1999). *The Invisible Computer: Why Good Products Can Fail, the Personal Computer Is So Complex, and Information Appliances Are the Solution*. MIT Press.

O'Brien, J., Rodden, T., Rouncefield, M., & Hughes, J. (1999). At home with the technology: an ethnographic study of a set-top-box trial. *ACM Transactions on Computer-Human Interaction (TOCHI), 6,* 282-308.

O'Hara, K., Mitchell, A. S., & Vorbau, A. (2007). Consuming video on mobile devices. In *Proceedings of CHI'07* (pp. 857-866). ACM Press.

O'Neill, T. M. (2002). Quality of Experience and Quality of Service
For IP Video Conferencing.

Oelbaum, T., Diepold, K., & a, W. (2007). A generic method to increase prediction accuracy of visual quality metrics. In *Proceedings of the Picture Coding Symposium*.

Okada, K.-I., Maeda, F., Ichikawaa, Y., & Matsushita, Y. (1994). Multiparty videoconferencing at virtual social distance: MAJIC design. In *Proc.ACM conf.on Computer supported cooperative work* (pp. 385-393).

Owens, D. A. & Wolfe-Kelly, K. (1987). Near Work, Visual Fatigue, and Variations of Oculomotor Tonus. *Investigative Ophthalmology and Visual Science, 28,* 743-749.

Owlsley, C. J., Sekuler, R., & Siemensne, D. (1983). Contrast Sensitivity through adulthood. *Vision Research, 23,* 689-699.

Pappas, T. & Hinds, R. (1995). On Video and Audio Integration for Conferencing. In *Proceedings of SPIE - The International Society for Optical Engineering*.

Persson, P. (1998). Towards a Psychological Theory of Close-ups: Experiencing Intimicay and Threat. http://www.kinema.uwaterloo.ca/pers981.htm

Pine, J. & Gilmore, J. (1999). *The Experience Economy*. Boston, MA, USA: Harvard Business School Press.

Pinson, M. & Wolf, S. (2004). A New Standardized Method for Objectively Measuring Video Quality. *IEEE Transactions on Broadcasting, 50,* 312-322.

Pitts, K. & Hurst, N. (1989). How much do people prefer widescreen (16x9) to standard NTSC (4x3)? *IEEE Transactions on Consumer Electronics, 35,* 160-169.

Podolsky, M., Romer, C., & McCanne, S. (1998). Simulation of FEC-based error control for packet audio on the internet. In *Proceedings of INFOCOM'98*.

Porter, G., Troscianko, T., & Gilchrist, I. (2003). Memory deployment in visual search: insights from pupillometry. *Journal of Vision, 3(5)*.

Poynton, C. (2003). *Digital Video and HDTV Algorithms and Interfaces*. San Francisco, CA, USA: Morgan Kaufmann.

Prangl, M., Szkaliczki, T., & Hellwagner, H. (2007). A Framework for Utility-based Multimedia Adaptation. *IEEE Transactions on Circuits and Systems for Video Technology, 17*.

Ramachandran, V. S. & Anstis, S. M. (1986). The perception of apparent motion. *Scientific American, 254,* 102-109.

Reeves, B., Detenber, B. H., & Steuer, J. (1993). New Televisions: The Effects of Big Pictures and Big Sound On Viewer Responses to the Screen. In *Information Systems Division*.

Reeves, B., Lang, A., Kim, E., & Tartar, D. (1999). The effects of screen size and message content on attention and arousal. *Media Psychology, 1,* 49-67.

Reeves, B., Lombard, M., & Melwani, G. (1992). Faces on the screen: Pictures or natural experience? In *Paper presented to the Mass Communication Division of the International Communication Association*.

Reeves, B. & Nass, C. (1998). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. University of Chicago Press.

Repo, P., Hyvönen, K., Pantzar, M., & Timonen, P. (2004). Users Inventing Ways to Enjoy New Mobile Services - The Case of Watching Mobile Videos. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences*.

Ribchester, E. (1958). Discussion Before th Radio and Telecommunication Section, 19th February 1958. *Proc.Institution of Electrical Engineering, 105B,* 437.

Richardson, I. (2003). *H. 264 and MPEG-4 video compression*. Chichester, England: John Wiley & Sons Ltd.

Richardson, I. E. G. & Kannangara, C. S. (2004). Fast subjective video quality measurement with user feedback. *Electronics Letters, 40,* 799-801.

Robin, M. (2003). Revisiting Kell. http://broadcastengineering.com/infrastructure/broadcasting_revisiting_kell/

Roetting, M. (2001). *Parametersystematik der Augen- und Blickbewegungen für arbeitswissenschaftliche Untersuchungen*. Aachen: Shaker.

Rogers, E. M. (1995). *The diffusion of innovations*. (Fourth edition ed.) New York: Free Press.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift fuer die Sozialpsychologie, 9,* 222-245.

Rubin, A. M. (1981). An examination of television viewing motivations. *Communication Research, 9,* 141-165.

Sadashige, K. & Ando, S. (1984). Recent development in large screen video display equipment technology. *Television Image Quality*.

Sakamoto, K., Aoyama, S., Asahara, S., Yamashita, K., & Okada, A. (2008). Relationship between Viewing Distance and Visual Fatigue in Relation to Feeling of Involvement. In S.Lee (Ed.), *APCHI 2008* (pp. 232-239). Berlin / Heidelberg: Springer.

Sanders, M. S. & McCormick, E. J. (1993). *Human Factors in Engineering and Design*. (7th ed.) New York, NY: McGraw-Hill Inc.

Sasse, M. A. & Knoche, H. (2006). Quality in Context - an ecological approach to assessing QoS for mobile TV. In *Proceedings of 2nd ISCA/DEGA Tutorial & Research Workshop on Perceptual Quality of Systems*.

Schubin, M. (2003). A foreword by Mark Schubin. In *Digital Video and HDTV: Algorithms and Interfaces* (San Francisco, CA, USA: Morgan Kaufman Publishers.

Schubin, M. (2007). Is there a perfect aspect ratio? http://www.theschubinreport.com/archive/04AM07-theschubinreport.mp3

Schuurman, D., De Marez, L., Veevaete, P., & Evens, T. (2009). Content and context for mobile television: Integrating trial, expert and user findings. *Telematics and Informatics, 26,* 293-305.

Schwartz, S. (2004). *Visual Perception*. (3rd ed.) McGraw-Hill.

Scott, F. (1955). Three-Bar Target Modulation detectability. *The Photogrammetric Record Photo.Sci.Engng, 10,* 49-52.

Seager, W., Knoche, H., & Sasse, M. A. (2007). TV-centricity - Requirements gathering for triple play services. In *Interactive TV: A Shared Experience TICSP Adjunct Proceedings of EuroITV 2007* (pp. 274-278).

Seferidis, V., Ghanbari, M., & Pearson, D. E. (1992). Forgiveness effect in subjective assessment of packet video. *Electronics Letters, 28(1),* 2013-2014.

Selier, C. & Chuberre, N. (2005). Satellite Digital Multimedia Broadcasting (SDMB) System Presentation. In *Proceedings of 14th IST Mobile & Wireless Communications Summit*.

Seo, K., Ko, J., Ahn, I., & Kim, C. (2007). An Intelligent Display Scheme of Soccer Video On Mobile Devices. *IEEE transactions on circuits and systems for video technology (CSVT), 17,* 1395-1401.

Serco (2006). Usability guidelines for Mobile TV design. Internet Available: http://www.serco.com/Images/Mobile%20TV%20guidelines_tcm3-13804.pdf

Seyler, A. J. & Budrikis, Z. L. (1964). Detail Perception after Scene Changes in Television Image Presentations. *IEEE Transactions on Information Theory, 11,* 31-42.

Shackle, B. (1990). Human factors and usability. In J.Preece & L. Keller (Eds.), *Human-Computer Interaction* (Prentice-Hall.

Shannon, C. E. (1949). Communication in the presence of noise. *Proc.Institute of Radio Engineers, 37,* 10-21.

Shenker, S. (1995). Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications, 13,* 1176-1188.

Shneiderman, B. (2003). *Leonardo's Laptop: Human Needs and the New Computing Technologies*. The MIT Press.

Siller, M. & Woods, J. (2003). Improving Quality of Experience for Multimedia Services by QoS Arbitration on a QoE Framework. In *IEEE PV 2003 Proceedings*.

Sinha, A. & Agarwal, G. (2005). A method of dynamic cropping and resizing of video frames in DVB-H to Mobile. In *GPSx 2005*.

Smith, S. L. (1979). Letter Size and Legibility. *Human Factors, 21,* 661-670.

Södergård, C. (2003). *Mobile television - technology and user experiences Report on the Mobile-TV project* (Rep. No. P506). VTT Information Technology.

Söhne, T., Flamm, P., Alrutz, H., Köhne, H., & Keller, S. (1998). A Video Backend for Multimedia TV-Sets. *IEEE Transactions on Consumer Electronics, 44,* 704-711.

Song, S., Won, Y., & Song, I. (2004). Empirical Study of User Perception Behavior for Mobile Streaming. In *Proceedings of the tenth ACM international conference on Multimedia* (pp. 327-330). New York, NY, USA: ACM Press.

Sporer, T. (1996). Evaluating Small Impairments with the Mean Opinion Scale - Reliable or Just a Guess? In *101st Audio Engineering Society Convention*.

Sproson, W. N. (1958). Discussion Before th Radio and Telecommunication Section, 19th February 1958. *Proc.Institution of Electrical Engineering, 105B,* 437.

Stanger, L. (2006). Submission to G-2.1.6 Progress Report of Task Force to define a Unit of Measure and Means of Calibration for Video Quality Analysis. http://grouper.ieee.org/groups/videocomp/lsrpt3d1.html

Steedman, W. C. & Baker, C. A. (1960). Target Size and Visual Recognition. *Human Factors, 2,* 121-127.

Steinman, R. M., Pizlo, Z., & Pizlo, F. J. (2000). Phi is not beta, and why Wertheimer's discovery launched the Gestalt revolution. *Vision Research, 40,* 2257-2264.

Steinmetz, R. (1996). Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communications, 14*.

Stewart, T. (2008). Usability or user experience - what's the difference? http://www.system-concepts.com/articles/usability%20%26%20hci/usability%20or%20user%20experience%20%11%20what%27s%20the%20difference?/

Storms, R. L. & Zyda, M. J. (2000). Interactions in Perceived Quality of Auditory-Visual Displays. *Presence, 9,* 557-580.

Strategy Analytics (2006). *TV Phones: Integration and Power Improvements Needed to Reach 100 Million Sales*.

Sugama, Y., Yoshida, T., Hamamoto, T., Hangai, S., Seng, B. C., & Kato, S. (2005). A Comparison of Subjective Picture Quality with Objective Measure Using Subjective Spatial Frequency. In *Proc.of ICME* (pp. 1262-1265).

Sylvers, E. (2007). Italia hails growth of its mobile TV service. http://www.iht.com/articles/2006/07/20/technology/italia.php

Tamminen, S., Oulasvirta, A., Toiskallio, K., & Kankainen, A. (2004). Understanding mobile contexts. *Special Issue of Journal of Personal and Ubiquitous Computing, 8,* 143.

Tang, J. C. & Isaacs, E. A. (1993). Why Do Users like Video? Studies of Multimedia-Supported Collaboration. *Computer Supported Cooperative Work: An International Journal, 1,* 163-196.

Tanton, N. E. & Stone, M. A. (1989). *HDTV Displays* (Rep. No. BBC RD 1989/9 PH-295). BBC Research Department, Engineering Division.

Taylor, A. & Harper, R. (2002). Switching on to switch off: An analysis of routine TV watching habits and their implications for electronic programme guide design. *usableiTV, 1,* 7-13.

Thang, T. C., Kang, J. W., & Ro, Y. M. (2007). Graph-Based Perceptual Quality Model for Audiovisual Contents. In *Proceedings of IEEE Multimedia and Expo, 2007* (pp. 312-315).

Thompson, F. T. (1957). Television line structure suppression. *SMPTE, 66,* 603-606.

Thompson, R. (1998). *Grammar of the shot*. Elsevier Focal Press.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34,* 273-286.

Tinker, M. (1963). *Legibility of Print*. Ames, Iowa: Iowa State University Press.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. Wiley & Sons.

Väänänen-Vainio-Mattila, K. & Ruuska, S. (2000). Designing mobile phones and communicators for consumer's needs at Nokia. In E.Bergman (Ed.), *Information appliances and beyond. Interaction design for consumer products* (pp. 169-204).

Vatakis, A. & Spence, C. (2006). Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. *Neuroscience Letters (submitted)*.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27,* 425-478.

Voldhaug, J. E., Johansen, S., & Perkis, A. (2005). Automatic Football Video Highlights Extraction. In *NORSIG-05*.

VQEG (2000). *Final Report from the video quality experts group on the validation of objective models of video quality assessment* (Rep. No. http://www.vqeg.org).

Vyas, D. & van der Veer, G. C. (2006). Experience as meaning: Some Underlying Concepts and Implications for Design. In *Proceedings of the 13th Eurpoean conference on Cognitive ergonomics: trust and control in complex socio-technical systems* (pp. 81-91). New York, NY, USA: ACM.

W3C-Recommendation (1998). Synchronized Multimedia Integration Language SMIL 1.0 Specification. http://www.w3.org/TR/1998/REC-smil-19980615

Wang, D., Speranza, F., Vincent, A., Martin, T., & Blanchfield, P. (2003). Towards Optimal Rate Control: A Study of the Impact of Spatial Resolution, Frame Rate and Quantization on Subjective Quality and Bitrate. In *Visual Communications a Image Processing*.

Wang, Z., Bovik, A. C., & Lu, L. (2002). Why is Image Quality Assessment So Difficult? In *IEEE International Conference on Acoustics, Speech, & Signal Processing*.

Ware, C. (2000). *Information Visualization*. Morgan Kaufmann Publishers.

Watson, A. (1996). Evaluating audio and video quality in low-cost multimedia conferencing systems. *Interacting with Computers, 8,* 255-257.

Watson, A. & Sasse, M. A. (1997). Multimedia conferencing via multicast: determining the quality of service required by the end user. In *Proceedings of AVSPN '97* (pp. 189-194).

Watson, A. B., Hu, J., & McGowan III, J. F. (2001). DVQ: a digital video quality metric based on human vision. *Journal of Electronic Imaging, 10,* 20-29.

Watson, A. & Sasse, M. A. (1998). Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. In *Proceedings of ACM Multimedia '98*.

Weber, E. H. (1836). *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*. Leipzig, Germany: Köhler.

Weiser, M. (1991). The Computer for the Twenty-First Century. *Scientific American, 265,* 94-104.

Weiss, S. (2006). World Handset Forum 2006. http://www.handheldusability.com/

Westerink, J. H. & Roufs, J. A. (1989). Subjective Image Quality as a Function of Viewing Distance, Resolution, and Picture Size. *SMPTE Journal, 98,* 113-119.

Westheimer, G. (1992). Visual acuity. In W.M.Hart (Ed.), *Adler's Physiology of the Eye: Clinical Application* (9th ed., St. Louis, Mo: CV Mosby.

Wheeler, H. A. & Loughren, A. V. (1938). Fine Structure of Television Images. *Proceedings of the Institute of Radio Engineers, 26,* 540-575.

Wilson, G. (2006). *Relationship between Media Quality and User Cost in Multimedia Systems.* Unpublished thesis: University College London.

Wilson, G. & Sasse, M. A. (1999). Listen to your heart-rate: Counting the cost of media quality. In *Proc.International Workshop on Affect in Interactions* (pp. 16-20).

Wilson, G. & Sasse, M. A. (2004). From doing to being: getting closer to the user experience. *Interacting with Computers, 16*.

Wilson, J. C. (1938). Channel Width and Resolving Power in Television Systems. *Journal of Television Society, 2,* 397-429.

Winkler, S. (1999). Issues in Vision Modeling for Perceptual Video Quality Assessment. *Signal Processing, 78,* 231-252.

Winkler, S. (2009). On the properties of subjective ratings in video quality experiments. In *Quality of Multimedia Experience* (pp. 139-144).

Winkler, S. & Faller, C. (2005). Maximizing audiovisual quality at low bitrates. In *Proc.of Workshop on Video Processing and Quality Metrics*.

Winkler, S. & Faller, C. (2006). Perceived Audiovisual Quality of Low-Bitrate Multimedia Content. *IEEE Transactions on Multimedia, 8,* 973-980.

Wood, D. (2004). *High Definition for Europe - a progressive approach* (Rep. No. EBU Project Group B/TQE). EBU.

Yanqing, C., Chipchase, J., & Jung, Y. (2007). Personal TV: A Qualitative Study of Mobile TV Users. In *Proceedings of EuroITV 2007* (pp. 195-204). Springer.

Yoshida, J. (2006). Mobile TV missing the goal. EE Times Available: http://www.eetimes.com/issue/fp/showArticle.jhtml?articleID=188703339

Yu, Z., Wu, H. R., & Ferguson, T. (2002). The Influence of Viewing Distance on Subjective Impairment Assessment. *IEEE Transactions on Broadcasting, 48,* 331-336.

Zeithaml, V. A., Parasuraman, A., & Barry, L. L. (1990). *Delivering Quality Service: Balancing Customer Perceptions and Expectations*. The Free Press.

Zettl, H. (1973). *Sight sound motion*. Belmont, CA, USA: Wadsworth.

Zhao, M., Bosma, M., & de Haan, G. (2007). Making the best of legacy video on modern displays. *Journal of the Society for Information Display, 15,* 49-60.

Zillman, D. (1988). Mood Management: Using Entertainment to Full Advantage. In L.Donohew, H. E. Sypher, & E. T. Higgins (Eds.), *Communication, Social Cognition, and Affect* (pp. 147-172). Hillsdale: Erlbaum.

Appendix

A 1. Focus group picture material for service examples

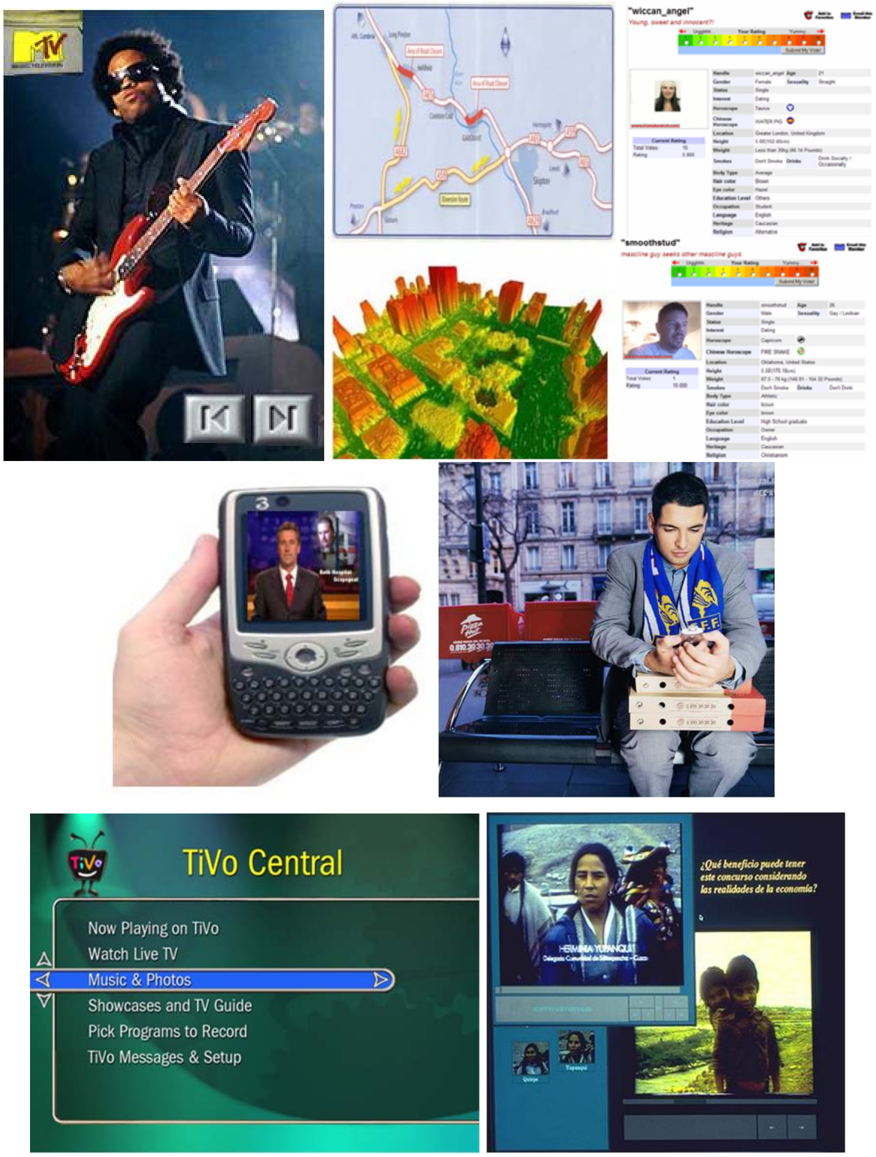


Figure 65: Focus group mobile content illustrations - music, disaster, dating, news, football, PVR, language

A 2. Debrief questionnaire from study 1

What was the reason when you pressed unacceptable?

Anything specific for

	General	Related to size
NEWS		
FOOTBALL		
MUSIC		
ANIMATION		

Interest

Would you be interested using a TV service like this on your mobile phone, at acceptable quality?

Which of the above content types would you like to watch (or any other)?

Pricing

How much would you be willing to pay for this if you could watch as much as you like?

£ per month

£ per day (for a day pass)

How much do you currently spend on your mobile phone per month?

What do you do when you are using public transport?
E.g. music, reading(books, newspapers (bought/free?)

A 3. Debrief questionnaire from study 2

What was the reason when you pressed unacceptable?

	General	Related to size
NEWS		

Did you have any difficulty with the text?

Interest

Would you be interested using a TV service like this on your mobile phone, at acceptable quality?

A 4. Debrief questionnaire from study 5a

What was the most important thing, for making your decision to watch the left or the right clip?

Once you had made your choice, did you look back, when?

What was the difference between the different sizes?

Which picture would you prefer left right (for the different sizes)

Did you perceive a difference in quality between the clips? if yes, did this affect your choice?

How often do you watch football? (never, rarely, once month, weekly)

Do you consider yourself a football fan?

Do you have a team you root for?

Would you watch football on a mobile TV, full games or only highlights?

A 5. Instructions for study 6

Welcome to my study on video clips on small screens. During the course of this experiment you will watch 16 very short video clips of varying content (football, news, music videos, and animation). You can watch these clips in six different sizes.

When a video starts please flip through all six of the sizes and decide, which produces the best visual experience for you. As you do this, please tell me, which size you would be most likely to watch and which sizes are NOT acceptable in terms of the viewing experience. It would be helpful if you could tell me why, as well.

After you have watched all 16 clips I will ask you some questions about the experiment, pay you for your time and you'll be good to go.

Feel free, of course, to leave the experiment at any time if you are not feeling well or for any other reason.
We are video taping the whole session to facilitate our analysis. The material will only be used for the analysis of the study. In case we would like to use your footage e.g. for a publication we would ask for your permission to do so.

If you have any questions please do not hesitate to ask me.

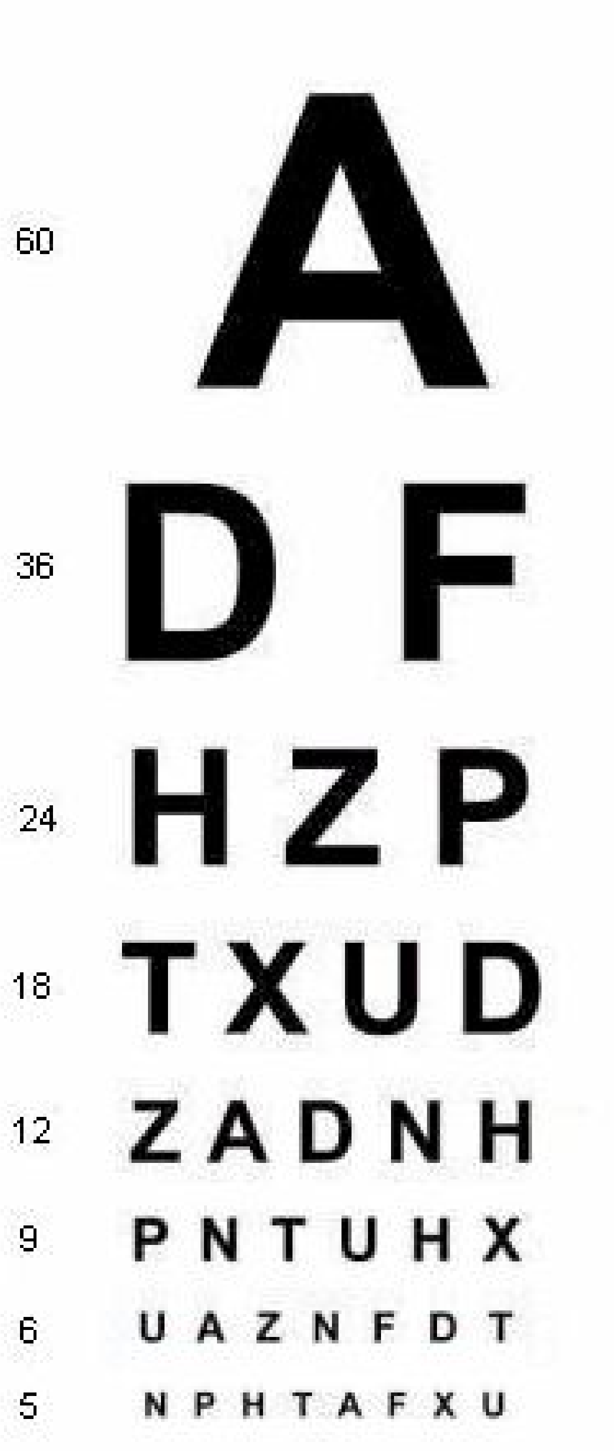
A 6. Rating sheet of study 6

Participant name: ____________ number: __ Stim order: _16_s2_

Which size do you find to be the best visual experience?
Which sizes do you find acceptable in terms of the visual experience for mobile TV?

1 file: _m4 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	2 file: _a3 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	3 file: _n2 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	4 file: _f1 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________
5 file: __f3 s2_ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	6 file: _a1 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	7 file: _m2 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	8 file: _n4 s2__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________
9 file: n1 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	10 file: f4 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	11 file: a2 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	12 file: m3 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________
13 file: f2 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	14 file: n3 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	15 file: a4 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	16 file: m1 s2 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________
x17 file: __f5c_ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	x18 file: _f6n__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	x19 file: _f5n__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________	x20 file: _f6c__ 1 □ ______________ 2 □ ______________ 3 □ ______________ 4 □ ______________ 5 □ ______________ 6 □ ______________

A 7. Snellen chart



When printed the letter A (line labelled 60) should be 44*mm* in height and the whole chart should be read by the participants from a distance of three meters.

A 8. Ishihara colour test