Journals
The Institution of Engineering and Technology

Volume 7 | ISSN 978-1-78561-129-2
2015-2016

The Best of IET and IBC 2015-2016

INSIDE Papers and articles on electronic media technology from IBC 2015-2016 presented with selected papers from the IET's flagship publication *Electronics Letters.*

Published by the Institution of **Engineering and Technology** www.**theiet**.org/journals

Contents

Selected content from IBC2015

Selected content from The IET

An electronic version of this publication and previous volumes can be found at www.theiet.org/ibc or by scanning the QR code

Introduction

Welcome to *The Best of IET and IBC* 2015-16. This is the seventh volume of an annual joint publication between the Institution of Engineering and Technology and IBC.

The IET is a formal member of the IBC's partnership board but, beyond this, it has a long-standing and close relationship with the organisation, through which they together encourage and promote professional excellence in the field of media technology. Nowhere is this relationship more strongly reflected than in the pages of this publication, which celebrates the very best technical media papers from this year's *IBC Proceedings* and the IET's flagship journal, *Electronics Letters.*

This year, our editorial takes a look at the exciting technology on show in IBC's Future Zone – an area of exhibition space where the world's most hi-tech media companies and research organisations proudly demonstrate their very latest concepts and experimental technologies. Here, you can not only *see* tomorrow's media but you have the opportunity to *experience* it personally, leaving impressions that will remain with you long after you have left Amsterdam.

We then present nine papers chosen as the best contributions to IBC2015 by the IBC Technical Papers Committee and the executive team of the IET Multimedia Communications Network. These include the overall winner of IBC's award for the Best Conference Paper, 'A Display-independent High Dynamic Range Television System' and papers representing other hot topics of 2015: UHDTV (Ultra High Definition Television), HEVC (High-Efficiency Video Coding), second screen applications, 4G broadcasting, MPEG DASH (Dynamic Adaptive Streaming over HTTP) technologies, hybrid content radio and some remarkable new results in on-screen subtitling.

We are also pleased to present personal interviews with individuals whose significant work appears in this volume. First, Andrew Cotton and Tim Borer of the BBC, authors of IBC2015's Best Conference Paper, who discuss their work on HDR. Find out where their inventive ideas come from, how they combine both psychology and engineering in their work and what the most memorable parts of their project have been.

We then interview IBC's Best Young Professionals: Raphaël Guénon and François Manciet. Both researchers in their mid-20s whose work has revealed some fascinating conclusions about second-screen media applications. Get a glimpse of their personal worlds. Find out: what motivates them to work in this area, whether they use second screens themselves and what they think about the future of immersive media.

From *Electronics Letters* this year we include a selection of media-related papers which have been published since IBC2014. *Electronics Letters* has a very broad scope, covering the whole range of electronics-related research and the papers chosen this year are those which we believe will have the greatest impact on media technology as well as the greatest potential for expanding service provision with existing infrastructures.

The IBC papers printed here represent the best of almost 300 synopses submitted to us this year by potential authors from across the world. Although this figure has remained remarkably constant for many years, the diversity of the topics covered certainly has not; it has continued to expand at an ever-bewildering rate. What began decades ago as terrestrial radio and television broadcast engineering, now encompasses every topic in the modern media technology world, culminating in such recent concepts as: the Internet of Things, Big Data, Long-term Evolution networks, Cloud applications, panoramic video, object-based media, audience engagement analytics, augmented soundscapes, human sensing and computational imaging.

We are extremely proud that so many media professionals continue to choose IBC for the publication of their technical work and as a forum for discussion with their fellow engineers and market strategists. This journal is a tribute to all those individuals who submitted synopses this year, whether successful or not. If you are inspired by the papers and stories presented here and would like to tell us about your own research or innovation, then please look out for our call for papers in January. And if your synopsis was not successful this year, then please try again - we work hard to accommodate as many papers and posters as we possibly can.

I hope that you enjoy reading this collection of the best papers as much as I and my committee of specialists and peer reviewers. We would like to convey our thanks to everyone involved in the creation of this year's volume, both at the IET and at IBC, and to extend our best wishes for a successful and exciting IBC2015.

Dr Nicolas Lodge
Chairman, IBC Technical Papers Committee

Who we are

IBC

IBC is committed to staging the world's best event for professionals involved in content creation, management and delivery for multimedia and entertainment services. IBC's key values are quality, efficiency, innovation, and respect for the industry it serves. IBC brings the industry together in a professional and supportive environment to learn, discuss and promote current and future developments that are shaping the media world through a highly respected peer-reviewed conference, a comprehensive exhibition, plus demonstrations of cutting edge and disruptive technologies. In particular, the IBC conference offers delegates an exciting range of events and networking opportunities, to stimulate new business and momentum in our industry. The IBC conference committee continues to craft an engaging programme in response to a strong message from the industry that this is an exciting period for revolutionary technologies and evolving business models.

The IET

The IET is one of the world's leading professional societies for the engineering and technology community, with more than 163,000 members in 127 countries and offices in Europe, North America and Asia-Pacific. It is also a publisher whose portfolio includes a suite of 27 internationally renowned peer-reviewed journals covering the entire spectrum of electronic and electrical engineering and technology. Many of the innovative products that find their way into the exhibition halls of IBC will have originated from research published in IET titles, with more than a third of the IET's journals covering topics relevant to the IBC community (e.g. IET: Image Processing; Computer Vision; Communications; Information Security; Microwave Antennas & Propagation; Optoelectronics, Circuits & Systems and Signal Processing). The IET Letters contained in this publication come from the IET's flagship journal, Electronics Letters, which embraces all aspects of electronic engineering and technology. Electronics Letters has a unique nature, combining a wide interdisciplinary readership with a short paper format and very rapid publication, produced fortnightly in print and online. Many authors choose to publish their preliminary results in Electronics Letters even before presenting their results at conference, because of the journal's reputation for quality and speed. In January 2010 Electronics Letters was given a fresh new look, bringing its readers even more information about the research through a colour news section that includes author interviews and feature articles expanding on selected work from each issue.

Working closely with the IET Journals team are the IET Communities team. The communities exist to act as a natural home for people who share a common interest in a topic area (regardless of geography); foster a community feeling of belonging and support dialogue between registrants, the IET and each other. Assisting each community is an executive team, made up of willing volunteers from that community who bring together their unique experience and expertise for the benefit of the group. Members of the Multimedia Communications Community executive team play an essential role in the creation of this publication in reviewing, suggesting and helping to select content. They contribute their industry perspectives and understanding to ensure a relevant and insightful publication for the broad community represented at IBC, showing the key part volunteers have to play in developing the reach and influence of the IET in its aim to share and advance knowledge throughout the global science, engineering and technology community.

Editorial

The Future Zone at IBC 2015

The Future Zone at IBC2015 is a unique gallery of 'hands-on' exhibits, brought together into a single exhibition area in the Park Foyer, next to Hall 8, in the midst of the RAI Centre. It features the very latest ideas, developments and disruptive technologies in our industry. A visit here is a glimpse into the future of content creation and delivery ... for our home, in our car and at work.

In selecting exhibits for the Future Zone, we select the latest cutting-edge projects and prototypes from R&D Labs and companies of all sizes around the world – from large multi-national organisations to the smallest start-ups. We look for technologies that are exciting and disruptive, and for exhibits where visitors can try out the concepts for themselves. We use four over-arching guidelines:

- the exhibit must be thought-provoking, innovative and valid
- it must not be a product that is already on the market - products can be shown on the company stands in the other Exhibition Halls
- wherever possible, it should consist of a practical demonstration, with hands-on interaction
- the 'quality of experience' for visitors to the stands should be measurable on the Richter Scale (!), and inspire the visitors to learn more.

The Future Zone at IBC2015 is divided into several themed areas:

- ***Future Zone Showcase – Beyond Reality***

  Here, visitors can immerse themselves in the worlds of augmented reality (AR) and virtual reality (VR), experience the intensity of 360° content and news footage, and discover new sensations with synthetic touch (haptics) and 'four dimensional' displays. These original and interactive demonstrations will give visitors hands-on contact with some truly amazing inventions, transporting maybe even the most doubtful customer to another world!
  The Showcase exhibits include:
  - VR 'technology' companies demonstrating hands on technology solutions
  - VR ' content' companies providing immersive experiences in special viewing booths
  - AR and immersive and interactive 'story-telling'.

- ***International Innovations***

  Ground-breaking technologies that are changing the way consumers around the world are accessing and engaging with new content have been brought together into this exciting international arena. Demonstrations by the major broadcasters and research organisations from the USA, Europe, China, Korea and Japan will enable visitors to understand the convergence, competition and the new technologies that are appearing in our marketplaces:
  - 8k UHDTV on 85" and 13" displays
  - 4K/8K UHDTV over 1Gbps IP networks
  - converged broadband/broadcast 4K UHDTV services
  - 2D images, depth enhanced to 3D
  - immersive telepresence and audio lossless coding
  - the IP studio – an immersive VR tour – and more!

- ***Poster Presentations***

  We are privileged to be able to exhibit Poster presentations in the Future Zone from university and industry researchers in countries across the globe, including Russia, Germany, Korea, Netherlands, Poland, UK, USA and Australia.

  Their topics range from embryonic ideas to real implementations with some fascinating results. The Posters will be exhibited by their authors throughout Friday and Saturday.

IBC Posters are normally exhibited as two carefully-crafted 'A0' display boards hung at eye-height. However, this year, we are experimenting with some 'Digital Poster' display techniques; and have worked with some of our authors to produce short PowerPoint presentations of their work for display on large flat screens in the Poster area. We look forward to visitor feedback on this experiment.

- ***Perfect Pixel Projects.***

What can the new disruptive broadcast technologies that are being developed, really achieve in terms of picture quality and the consumer's quality of experience. Can the international projects and collaborations exhibiting here, and which are bringing together the world's leading edge researchers and implementers, really give us 'perfect pixels' on every platform? See and experience here:

- new developments for hybrid distribution of TV programmes and services
- the 'Wall of Moments' bringing together diverse sources of content
- pro-active 2nd screen devices for hybrid TV
- computer vision techniques for post-production workflows
- HDR 'video tone' mapping, plus glasses-free autostereoscopic 3D display .

The Future Zone is open to all, throughout the IBC Show; and we draw your attention in particular to the IET Reception, held in the Zone at 16:00 on Friday the 11th September, highlighting the Future Zone exhibits and launching the 2015 edition of the 'Best of IET and IBC' Journal, containing our organisations' best technical papers and articles of the year. At this event also, IET President-elect, Naomi Climer, will talk about her experiences working in the creative environment of America's Silicon Valley, and how new funding methods have stimulated the development of novel ideas and innovations. This is a great opportunity to network with the 'movers and shakers' in our industry, plus there will be complimentary refreshments from the IET to help nourish your stimulated mind.

This year's Future Zone pushes the boundaries of reality ... come and experience our future world.

Selected content from IBC2015

A “display independent” high dynamic range television system

T. Borer A. Cotton

BBC R&D, 56 Wood Lane, London W12 7SB, UK

Abstract: High Dynamic Range (HDR) television has captured the imagination of the broadcast and movie industries. This paper presents an overview of the BBC’s “Hybrid Log-Gamma” solution, designed to meet the requirements of high dynamic range television. The signal is “display independent” and requires no complex “mastering metadata”. In addition to providing high quality dynamic range (HDR) pictures it also delivers a high quality “compatible” image to legacy standard dynamic range (SDR) screens and can be mixed, re-sized and compressed using standard tools and equipment. The technical requirements for a high quality HDR television system are presented. Quantisation effects (or “banding”) are analysed theoretically and confirmed experimentally. It is shown that quantisation effects are comparable or below competing HDR solutions. The psychovisual effects of presenting images on emissive displays in dim environments, “system gamma”, is shown experimentally and the analysis informs the design of this HDR system.

Introduction

With improvements in technology, television with greater impact, more “presence”, deeper “immersion”, a “wow factor”, or, in short, better pictures, are now possible. Ultra high definition, UHD, is not just about more pixels, it has the potential to deliver wider colour gamut (WCG), higher frame rates (HFR), and higher dynamic range (HDR). Of these, perhaps high dynamic range offers the greatest improvement and the costs of upgrading to HDR can be relatively low for both production and distribution. High dynamic range offers unmistakeably better pictures, across the living room, on smaller displays, and even to those with less than perfect vision. Potentially HDR may be produced using mostly installed, legacy, standard dynamic range (SDR) infrastructure and distributed over largely unchanged distribution networks. No wonder that, even before the standards have been finalised, some movie studios are already talking about creating movies in HDR, Ultra HD for home viewing.

This paper describes the signal processing technology required for high dynamic range in the television production and distribution chains. It describes how one solution, the “hybrid log-gamma” approach, provides a “display independent” signal that can produce high quality images, which maintains the director’s artistic intent on a wide range of displays in diverse viewing environments. So, for example, precisely the same signal may be viewed in a controlled production suite, a home cinema, an ordinary living room or on a laptop or mobile device. Furthermore, the signal may be displayed on a conventional standard dynamic range display to provide a high quality “compatible” image. The log-gamma HDR signal may be mixed, re-sized, compressed, and generally “produced”, using conventional tools and equipment. The only specifically high dynamic range equipment needed is cameras and displays for quality monitoring (signal monitoring may continue to use SDR displays). No complex mastering metadata is required. Conventional end user distribution techniques may be used (although a 10 bit signal is required). No layered or multichannel codecs are required. Only a single signal is required for both SDR and HDR displays and expensive multiple “grades” (for both HDR and SDR) are not necessary.

The paper continues by discussing the meaning of high dynamic range. To understand HDR production and display we need to look at the television signal chain, which is discussed next. This then allows us to consider the design of the camera transfer characteristic (the opto-electronic transfer function (OETF)). Next we discuss an important psychovisual aspect of HDR TV, the “system gamma”. Whilst this effect has long been known in television, movies and the academic literature, it assumes an enhanced importance for HDR. Based on an understanding of system gamma we discuss the design of the electro-optic transfer function (EOTF) in the display, and how this can allow the display of high quality pictures on a diverse range of displays. Once the EOTF is defined we can analyse the likely effect of quantisation and the performance, in terms of dynamic range, that may be expected from the system. This is compared to alternative HDR proposals and the theoretical analysis is compared to experimental results. The paper ends with some concluding remarks.

Dynamic range

Dynamic range is the ratio between the whitest whites and blackest blacks in an image. For example printed images have a dynamic range of less than 100:1 (because it is difficult to make a black ink that reflects less than 1% of incident light). Dynamic range is often measured in "stops", which is the logarithm (base 2) of the ratio. Printed images have less than 7 stops of dynamic range. Standard dynamic range consumer television (8 bit video, e.g. DVD, SD and HD DVB) only supports about 6 stops of dynamic range, as discussed below. Professional SDR video (10 bits) supports about 10 stops. But the human eye can see up to about 14 stops (1) of dynamic range in a single image. Higher dynamic range results in an experience closer to reality, and hence of greater impact or "immersion". Furthermore higher dynamic range also increases the subjective sharpness of images and so provides a double benefit.

Some debate has confused high dynamic range with high brightness. The two are not the same. You can have high dynamic range in a dark movie environment, with pictures of only 48cd/m2. Alternatively you can have standard dynamic range on very bright screens of hundreds, or even thousands, of cd/m2. What high brightness does allow, is to see high dynamic range without needing a very dark viewing environment.

It might be thought that higher dynamic range could be achieved by simply making displays brighter. But this is analogous to suggesting that you can increase the dynamic range of audio by turning up the volume. With audio, turning up the volume merely emphasises the noise. The same is true for video. With video the "noise" is quantisation noise, where the steps between quantisation levels become clearly visible. This is manifest as "banding", "contouring, or "posterisation". An extreme example of banding is shown in Figure 1.

The useful dynamic range of video is determined by the ratio between adjacent quantisation levels. If the adjacent quantisation levels differ in luminance by less than 2% the difference is probably imperceptible in the image. This threshold of visibility increases at low luminance. The Schreiber curve[1] (shown later) is an approximation to the threshold of visibility.

In television systems the luminance represented by the signal is a non-linear function of the signal value. Conventionally this is a gamma curve, exemplified by ITU-R Rec 1886. For example, the displayed luminance L may be the signal V raised to the power gamma, $L = V^g$. The ratio between adjacent quantisation levels, also known as the Weber fraction, is given mathematically by:

$$\text{Weber fraction} = \frac{1}{\mathrm{N} \cdot \mathrm{V}} \frac{dL}{dV} = \frac{g}{N} \frac{1}{L^{1/g}}$$

Where N is the number of quantisation levels (220 for 8 bit video, 876 for 10 bit), and the algebraic form is given for a conventional gamma curve. Using this equation we may find the luminance corresponding to the threshold at which banding is visible, which then gives the usable dynamic range. For conventional, 8 bit SDR video, with gamma 2.4 and a 5% threshold (allowing for a dim display), this yields a dynamic range of only 5.27 stops. This is not a high dynamic range, a bit less than a photographic print, although it can be extended by using techniques such as dither. Overall the conventional display gamma curve is not adequate for HDR reproduction and a different non-linearity is required.

Television signal chain

The television signal chain, shown in Figure 1, intentionally includes signal non-linearities in both cameras and displays. The camera non-linearity is known as the opto-electronic transfer function (OETF) and the display non-linearity is known as the electro-optic transfer function (EOTF). Colloquially and confusingly in conventional television both transfer functions are known as "gamma curves". As is well known the EOTF is not the inverse of the OETF, so overall the signal chain has a non-linear (or "system") opto-optic transfer function (OOTF). The OOTF compensates for the psycho-visual effects viewing pictures on an emissive display in dim or dark surroundings, and is sometimes known as "rendering intent" (2). This is discussed in more detail below.

Originally the OETF, in combination with the CRT display EOTF, was designed to make the effects of camera noise more uniform at different brightnesses. In digital systems the non-linearities help to minimise the visibility of quantisation (or "banding"). But, by modifying these non-linearities, it is possible to further reduce the effects of banding and so increase dynamic range.

In this conventional model of the television chain, the relative luminance in the scene is captured by the camera and encoded in the signal. The light from the scene defines the signal. The EOTF then renders the light from scene so that, *subjectively*, it appears the same as reality. Historically, with CRT displays, display brightness was fairly consistent because CRTs simply could not be made very bright. This allowed a single EOTF to be used to render the signal for all CRT displays. With the availability of a plethora of bright display technologies different EOTFs are needed to ensure that pictures look *subjectively* the same on displays of different brightness. The approach described in this paper allows the signal to be rendered on any display (OLED,

[1] Schreiber measured the threshold of contrast visibility experimentally (3). Schreiber's results are broadly consistent with experiments on video quantization reported by Moore (4), and also with the DICOM model (5), which itself is derived from the Barten model (6).

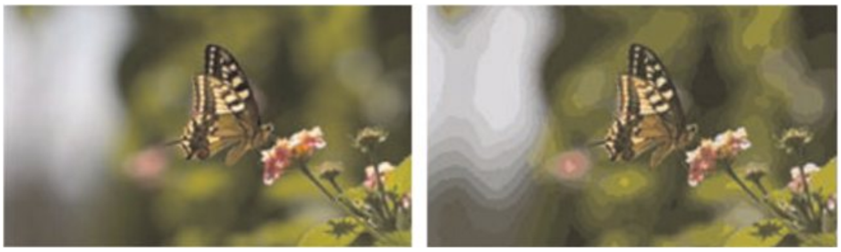


Figure 1 *Image Quantisation, left original, right extreme banding*

LCD, local backlight dimming, or quantum dot), preserving the director's artistic intent, without the need for metadata and without needing to re-grade for different displays. That is, the hybrid log-gamma approach defines a signal that is independent of the display.

The hybrid log-gamma opto-electronic transfer function (OETF)

In the brighter parts and highlights of an image the threshold for perceiving quantisation is approximately constant (know as Weber's law). This implies a logarithmic OETF would provide the maximum dynamic range for a given bit depth. Proprietary logarithmic OETFs, such as S-Log, Panalog and Log C are, indeed, widely used. But in the low lights it becomes increasingly difficult to perceive banding. That is, the threshold of visibility for banding becomes higher as the image gets darker. This is known as the De Vries-Rose law. The conventional gamma OETF comes close to matching the De Vries-Rose law, which is perhaps not coincidental since gamma curves were designed for dim CRT displays. So an ideal OETF would, perhaps, be logarithmic in the high tones and a gamma law in the low lights, which is essentially the form of the hybrid log-gamma OETF.

The dynamic range of modern video cameras is considerably greater than can be conveyed by a video signal using a conventional gamma curve (i.e. ITU-Rec 709). In order to exploit their full dynamic range conventional video cameras use a "knee" characteristic to extend the dynamic range of the signal. The knee characteristic compresses the image highlights to prevent the signal from clipping or being "blown out" (overexposed). A similar effect is also a characteristic of analogue film used in traditional movie cameras. When a hybrid gamma HDR video signal is displayed on a conventional SDR display, the effect is similar to the use of a digital camera with a knee or using film. It is not surprising therefore, that the hybrid gamma video signal is highly compatible with conventional SDR displays, because what you see is very similar to the signal from an SDR camera. Indeed the knee characteristic of the hybrid gamma characteristic, defined below, is conservative, providing only 300% overload.

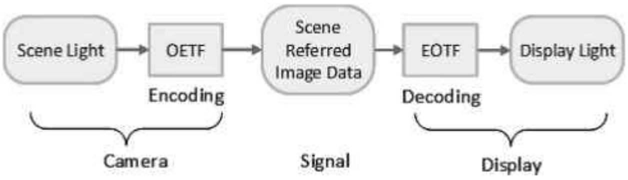


Figure 2 *Television signal chain*

A hybrid gamma signal is defined as:

OETF

$$E' = \begin{cases} r\sqrt{E} & 0 \leq E \leq 1 \\ a\ln(E-b)+c & 1 < E \end{cases}$$

where E is proportional to the light intensity detected in a camera colour channel (R, G, or B), normalized by the reference white level. E' is the non-linear, or "gamma corrected" signal, where the non-linearity is applied separately to each colour channel. The reference value of E' is 0.5, denoted "r", and corresponds to reference white level. Constants, a= 0.17883277, b= 0.28466892, c= 0.55991073, are defined so that the signal value is unity for a (relative) luminance of 12.0.

The hybrid log-gamma OETF is shown below alongside the conventional SDR gamma curve and a knee characteristic. Note that the horizontal axis for the hybrid log-gamma curve, as defined above, has been scaled to emphasise compatibility with the conventional SDR gamma curve. Furthermore, because the hybrid log-gamma

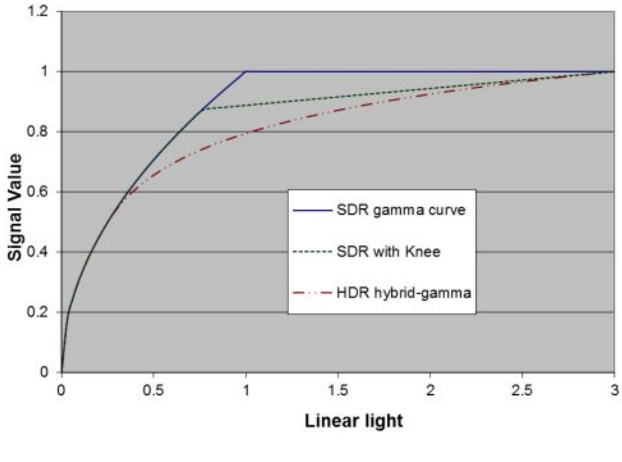


Figure 3 *Hybrid log-gamma and SDR OETFs*

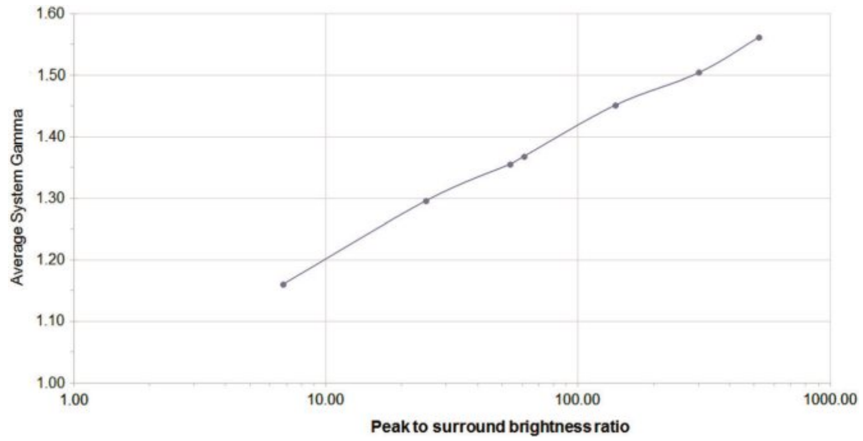


Figure 4 *Preferred system gamma versus normalised display brightness*

signal only describes the light representing the scene, it is independent of the display. Consequently, with a suitable EOTF, it may be used with any display.

System gamma and the opto-optic transfer function (OOTF)

As is well known, and noted above, the light out of a television display is not proportional to the light detected by the camera. The overall system non-linearity, or "rendering intent" (2) is defined by the opto-optic transfer function, or OOTF. Rendering intent is needed to compensate for the psychovisual effects of watching an emissive screen in a dark or dim environment, which affects the adaptation state (and hence the sensitivity) of the eye. Without rendering intent, pictures would look too bright or "washed out". Traditionally movies were, and often still are, shot on negative film with a gamma of about 0.6. They were then displayed from a print with a gamma of between 2.6 and 3.0; this gives movies a system gamma of between 1.6 and 1.8, which is needed because of the dark viewing environment. Conventional SDR television has an OOTF which is also a gamma curve with a system gamma of 1.2. But, for HDR, the brightness of displays and backgrounds will vary widely, and the system gamma will need to vary accordingly.

In order to determine the necessary system gamma we conducted experiments viewing images with different gammas at different luminances (and with a fixed background luminance). The pictures were derived from HDR linear light images selected from Mark Fairchild's HDR Photographic Survey. A reference display (Dolby PRM4220) and a test display (SIM2) were placed about 1 metre apart, in a controlled viewing environment (room illumination 10Lux, D65). The reference image, shown at 600cd/m^2 on the reference display, was chosen by participants from 9 images with different gammas (1.0 to 2.4). This allowed them to choose the artistic effect they preferred. The participants then choose a picture on the test display, from 9 different gammas (1.0 to 2.4) that best matched the reference image. The test images were shown at different luminances on the test display. The results, illustrated below, provide an estimate of the preferred system gamma, (excluding artistic preferences), at a range of display brightnesses from 68 to 5200cd/m^2. Whilst only a small number of participants were involved, and further experimental results would be most welcome, the results are quite consistent and provide a good guide to the necessary system gamma for different display brightness relative to background luminance.

These empirical results may be approximated by the following formula for system gamma, where Y represents luminance.

$$g = 1 + \frac{1}{5}\log_{10}\left(\frac{Y_{peak}}{Y_{surround}}\right)$$

The results clearly show that the end-to-end system gamma of the HDR TV system has to be adjusted to accommodate displays of differing peak luminance. They suggest that, with a background luminance of 10cd/m^2, an OLED display of around 1000 cd/m^2 would require a system gamma of around 1.4, whilst a brighter LCD of a few thousand cd/m^2 would require a system gamma closer 1.5. Whilst these variations in system gamma appear small, they have a significant impact on the subjective appearance of an image.

The hybrid log-gamma electro-optic transfer function (EOTF)

In order to specify the complete television system we need an EOTF as well as the OETF defined above. This maps the relative light representing the scene to the light emitted from the display. The EOTF should perform this mapping 1) whilst preserving the artistic intent of the programme maker (and providing a suitable rendering intent), 2) allowing for the dynamic range of the display from black level to peak white, and 3) minimising quantisation artefacts. The EOTF defined below is similar to the

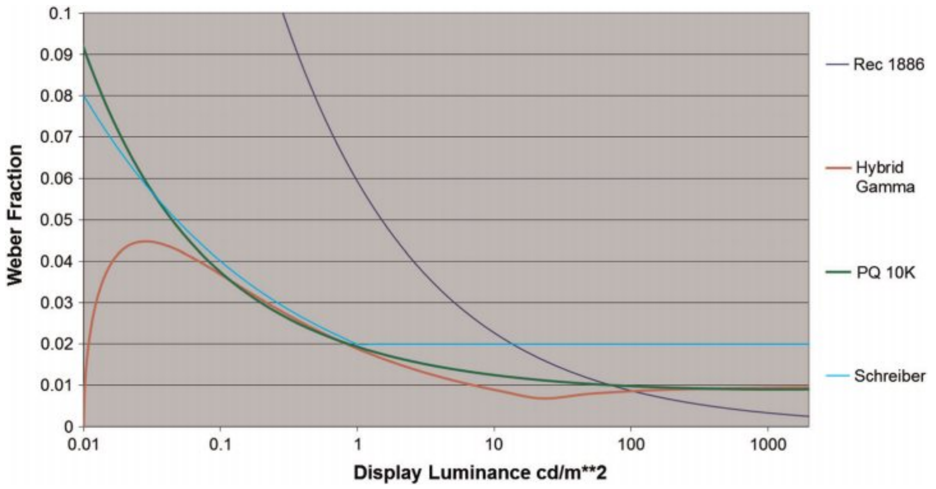


Figure 5 *Weber fractions versus display luminance*

conventional display gamma curve, thereby maximising backward compatibility, whilst also meeting the three preceding requirements;

EOTF

$$Y_d = aY_s^g + b$$

where $\mathrm{Y_d}$ is the luminance of a pixel presented on the display, $\mathrm{Y_s}$ is the relative luminance representing the scene for that pixel, and *g* is the system gamma discussed above. Parameters *a*, and *b* correspond to similar parameters in ITU-R Rec 1886, which are the traditional "contrast" and "brightness" controls respectively. They determine the peak displayed luminance and the minimum luminance, i.e. the black level[2].

The EOTF maps the linear scene luminance, $\mathrm{Y_s}$, to the linear display luminance, $\mathrm{Y_d}$. This differs from current practice for SDR, which applies the EOTF to each colour component independently. But applying the EOTF to each component changes the saturation, and to a lesser extent hue, of the picture. Since the EOTF needs to change with the display it must be applied to luminance to avoid inconsistent colours.

Scene luminance Ys may be recovered from the signal by first applying the inverse of the OETF to each colour component R', G', and B' to yield the linear colour components R, G and B. With the same nomenclature as the OETF;

Inverse OETF

$$E = \begin{cases} (E'/r)^2 & 0 \le E' \le r \\ \exp((E' - c)/a) + b & r < E' \end{cases}$$

From the linear colour components the scene luminance may be derived as follows (assuming ITU-R Rec 2020 colorimetry);

Scene luminance

$$Y_s = \frac{0.2627\ R + 0.6780\ G + 0.0593\ N}{12}$$

Note that the factor of 12 in the denominator is because the signal normalisation for the OETF yields a maximum value of 12 for each linear colour component, rather than the more conventional value of 1.

Having determined the linear scene luminance the displayed luminance may be derived from the EOTF, where parameters *a*, *b*, and *g* depend on the display and the viewing environment. Given the displayed luminance we still need to determine the individual $\mathrm{R_d}$, $\mathrm{G_d}$, and $\mathrm{B_d}$ values that should be displayed for each pixel. We obtain these simply by scaling the linear scene colour components as follows:

Displayed colour components

$$\begin{aligned} R_d &= R \times ((Y_d - b)/12Y_s) + b \\ G_d &= G \times ((Y_d - b)/12Y_s) + b \\ B_d &= B \times ((Y_d - b)/12Y_s) + b \end{aligned}$$

where $\mathrm{R_d}$, $\mathrm{G_d}$, $\mathrm{B_d}$, are the luminances presented on the display.

Minimising the visibility of quantisation, or "banding" is an important aspect of the EOTF. We can estimate its visibility by calculating the weber fraction and comparing it to the Schreiber limit, as discussed above. Doing so and plotting the results yields the flowing graph for Weber fraction versus displayed luminance. The EOTF may be used for displays with different peak luminance and black levels. This example assumes a 10 bit signal, with peak

[2]
$$a = L_P - L_B \quad b = L_B$$
where L_P is the displayed luminance for peak white ($Y_s = 1.0$), and L_B is the displayed luminance for black ($Y_s = 0.0$).

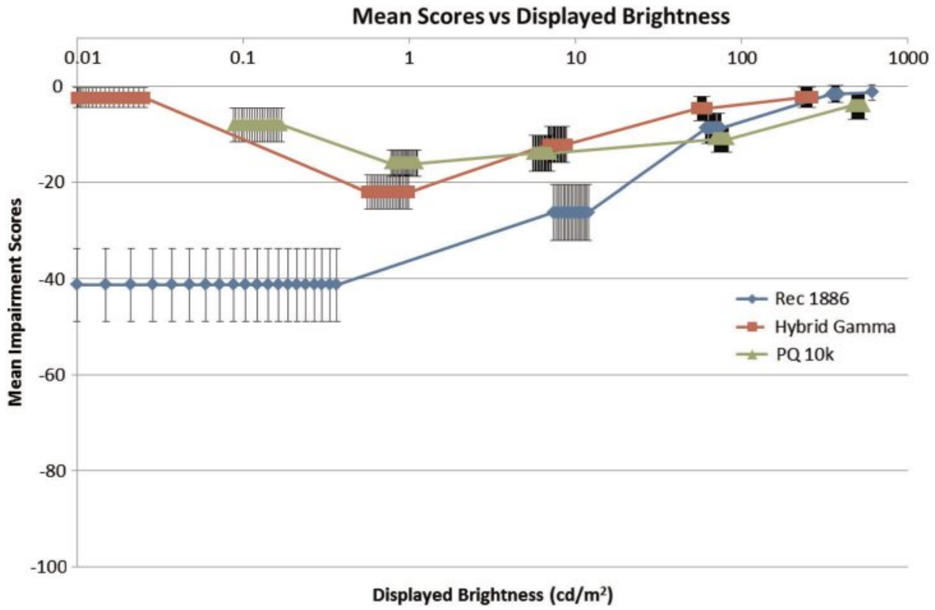


Figure 6 *Banding impairment versus displayed brightness*

white of 2000 cd/m^2, a black level of 0.01 cd/m^2, and a system gamma of 1.5. This corresponds to the use of bright displays for programme monitoring and grading, viewed in a "dark" environment relative to the brightness of the display. It represents a dynamic range of 200,000:1 or 17.6 stops, which is more than the dynamic range the eye can perceive in a single image.

For comparison this graph also includes the Schreiber limit, the conventional SDR gamma curve (Rec 1886), and an alternative HDR, a perceptual quantisation curve (PQ 10K) defined in SMPTE ST 2084. Banding is likely to be visible when the Weber fraction for an EOTF is above the Schreiber limit. With a 10 bit signal this indicates that for the hybrid gamma EOTF banding will, at worst, be at the threshold of visibility across the whole luminance range and similar to or below that of the PQ curve. It also shows that the conventional gamma curve is not adequate, with banding expected to be visible below 20 cd/m^2. Note that this analysis is for a 10 bit signal. With a 12 bit signal, which has been proposed for HDR production, Weber fractions would be much lower and banding would be significantly below the threshold of detectability across the whole luminance range.

To confirm the theoretical analysis we performed experiments on the comparative visibility of banding. Highly critical 10 bit, horizontal, "shallow ramps", with adjacent patches varying by 1 quantisation level, were compared to a continuous reference[3] using the ITU-R Rec 500 double stimulus impairment scale method. 33 subjects were tested. The test images were displayed on a Dolby PRM 4220 monitor configured (using a custom internal LUT) to emulate the low tones of a display with 2000cd/m^2 peak luminance and black level of 0.01cd/m^2. Each horizontal grey scale ramp occupied 25% of screen height, and included 20 adjacent grey levels each spanning 1/24th picture width. The experimental results are shown below. In these results -20 is one grade of impairment. The error bars indicate the 95% confidence intervals.

These results appear to closely corroborate the theoretical analysis. Both the hybrid gamma and ST 2084 are less than or about 1 grade of impairment, which is described as "imperceptible" or just "perceptible but not annoying" at their very worst. Furthermore the hybrid gamma EOTF shows marginally more banding than ST 2084 in the region of 1cd/m^2, and marginally less banding elsewhere, which is in line with the theoretical analysis. The impairment for conventional SDR gamma (ITU-R Rec 1886) rises from "imperceptible" to "perceptible but not annoying" below 20cd/m^2, up to "slightly annoying" in the region below 1cd/m^2, also in line with the theoretical analysis. Overall the hybrid gamma curve provides acceptable banding performance at 10 bits for highly critical material, equivalent to that of ST 2084, for a display with 17.6 stops of dynamic range. In practice it is highly unlikely that any banding will be visible on naturally occurring scenes.

Conclusions

This paper has presented the rationale and design for a HDR television system. The hybrid gamma approach can support a range of displays of different brightness, without metadata, and so is display independent. A 10 bit signal is substantially compatible with conventional SDR signals. Shown unprocessed on an SDR display the picture is of high quality and so may be used for signal monitoring.

[3] The reference was a carefully dithered 10 bit signal in which banding was undetectable.

This also means production can use existing SDR infrastructure, tools and equipment. Only quality monitoring requires a HDR display. No metadata is required, thereby simplifying the production chain. These features mean SDR production may be upgraded to HDR at relatively modest cost. Only a single signal is required for both SDR and HDR displays and expensive multiple "grades" (for both HDR and SDR) are not necessary. Consequently layered, or multichannel, coding for end users is unnecessary, thereby simplifying distribution and minimising cost. For a 2000cd/m^2 HDR display, with a black level of 0.01cd/m^2, i.e. 17.6 stops dynamic range, it is shown, both theoretically and experimentally, that quantisation artefacts ("banding") will not be visible on real pictures and that "banding" is comparable, or less, than the competing, more complex, ST 2084 HDR system. Finally note that HDR Rec 2020 signals may be formatted to look like conventional SDR Rec 709 signals, raising the possibility of conventional SDR media carrying HDR signals in a completely compatible way. This will be the subject of a future paper.

Acknowledgement

We would like to thank Manish Pindoria of BBC R&D for his work on this project, without which this paper would not have been possible.

References

[1] KUNKEL T., REINHARD E.: 'A reassessment of the simultaneous dynamic range of the human visual system', *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, July 2010 ISBN 978-1-4503-0248-7, pp. 17–24

[2] POYNTON C.: 'Digital Video and HD' (Morgan Kaufmann, 2 December 2012, 2nd edn.), ISBN-13: 978-0123919267

[3] SCHREIBER W.F.: 'Fundamentals of Electronic Imaging Systems' (Springer-Verlag, 1992, 3rd edn.), ISBN: 978-3-540-56018-0

[4] MOORE T.A.'Digital video: the number of bits per sample required for reference coding of luminance and colour-difference signals', BBC Research Department Report, 1974, BBC RD 1974/42

[5] DICOM: Digital Imaging and Communications in Medicine (DICOM), Part 14: Grayscale Standard Display Function (National Electrical Manufacturers Association, PS3.14, 2008)

[6] BARTEN P.G.J.: 'Formula for the contrast sensitivity of the human eye', *Proceedings SPIE-IS&T*, January 2004, Vol. 5294, pp. 231–238

Interview - Andrew Cotton & Tim Borer

1. Tell us a bit about yourself and what you do

AC: I am a Principal Technologist at BBC Research and Development and have a background in video compression and image processing. I co-ordinate the BBC's UHDTV standardisation activities, but in addition my team and I are responsible for maintaining the technical integrity of the BBC's acquisition, production, playout and IP distribution systems. I joined BBC R&D in 1987 after graduating with a BA in Engineering Science.

TB: I work as a Lead Engineer at BBC Research and Development, currently focusing on aspects of UHDTV such as high dynamic range and high frame rates. Prior to the BBC I designed professional broadcasting equipment, including motion compensated standards converters and compression equipment, for both Snell and Harris. Tim holds degrees in video processing, electronics and physics. He is a Chartered Engineer (MIET), a senior member of the IEEE and a member of the SMPTE. He is the inventor of 20 patents.

2. How would you describe HDR television to someone who has never seen it?

Combined answer:

Imagine a television programme you have watched which has really good pictures, perhaps a high quality natural history programme. Then imagine this with more detail and vibrancy, with sparkling highlights on water drops, bird's feathers and animal's whiskers, and with brighter colours. This is HDR television, in which the pictures feel more real and more like you are looking at the real world. HDR is probably what many people were hoping would be (but wasn't) the step change to UHD. With even only HD resolution the pictures are clearly higher quality, which can be appreciated by everyone, and with greater resolution from UHD they look even better.

Often the first time people see HDR it gives them goose bumps. The pictures come alive. It is like adding an extra dimension. The level to which it will improve the viewing experience for audiences is palpable.

3. What were the biggest challenges that you faced in undertaking this project?

Combined answer:

There were a number of large challenges we had to address, the first of which was the lack of understanding of how television actually works within our industry. The end-to-end workflows, production environments and operational constraints are quite different from digital cinema and Blu-ray, and not many people appreciate that.

Also trying to convince people th at the assumptions they had taken for granted with SDR television did not necessarily hold true for HDR and needed to be modified. Many of the issues, such as the role of gamma, have actually been well known for a long time but because they have remained unchanged they have been forgotten or misunderstood. This now probably remains our greatest challenge in getting acceptance of HDR.

4. Tell us how the novel ideas in your design approach arose.

AC: I was working in DVB and struggling to understand how broadcasters could launch UHD HDR services, whilst delivering something useful to the many millions of UHD SDR receivers that will already be in people's homes. Having worked on layered coding solutions for many years, we both knew that they were seldom (if ever) successful in the market. We needed something simpler, and Tim, with his expert knowledge, thought that this mad idea of combining a log-curve with the existing gamma-curve might just work. Moreover, it would not require Production Metadata, which is the bane of many a television engineer's life. Tim then did the tricky part of making the mathematics work.

TB: As always in a new project new ideas arrive through a combination of thinking and discussion with experts in the industry, reading the literature, conducting experiments and, above all, looking at pictures. As we learnt more this information fused together and resulted in a new understanding and novel ways to approach HDR TV.

5. Do you feel privileged to have had the rare opportunity to define a fundamental element of a new television standard?

Combined answer:

We certainly do feel privileged to have participated in the development of fundamentally new standards for TV. In the near future we hope to have set of standards addressing a new generation of television systems which embraces not just HDR but wider colour gamut and higher frame rates, which are all interrelated, as well as UHD. All engineers wish to see their ideas adopted in practice and we have been fortunate to work on a project that we hope will be widely adopted.

6. Your approach successfully draws upon some classical studies of the psychology of human vision, has this been a pleasing part of the work?

TB: It has been interesting, whilst researching the psychovisual aspects of HDR, to realise that many of the fundamental issues have been investigated and known for many decades. It has been fascinating to see how this academic work fits together and has been proved to be correct in the context of HDR.

7. How many alternative HDR systems are under consideration internationally and how do they differ from yours?

Combined answer:

There are effectively two approaches to HDR being considered by the industry for short to mid-term deployment. Our approach is “scene referenced” whilst the other approaches are “display referenced”. Fundamentally that means they require either explicit or implicit metadata to describe the mastering environment, whilst our system does not require any metadata. A number of other systems have been proposed for the longer term based on alternative colour spaces. But it’s not yet clear whether they will offer worthwhile advantages in either Production or Distribution.

Our approach is most similar to conventional TV, and even provides backward compatibility, whereas the alternative approaches are more focused on the approach of the movie industry. Both approaches can provide the highest quality HDR images. Our approach supports the needs of the television industry by allowing an evolutionary approach to HDR production and distribution, mostly using conventional broadcast equipment and infrastructure, as demonstrated at IBC. It allows production without requiring metadata and distribution, over internet, terrestrial and satellite, using a single compressed stream, as we do presently. It also supports presentation on a wide variety of displays such as LCDs with local dimming, OLED and quantum dot displays. This allows the market to produce the televisions consumers prefer, helped by the non-proprietary nature of our approach.

8. We can buy UHD TV sets now and trial transmissions are taking place. When do you think that we might be able to see HDR TV in the home?

Combined answer:

Several manufacturers already have HDR screens in their line-up, and we’re aware of two HDR streaming services available to subscribers with the right type of equipment. But DVB, amongst others, is targeting 2016 for the completion of its HDR broadcast specification. There are a number of standardisation and regulatory hurdles to be addressed before that can happen. And it usually takes about a year from the publication of those standards to suitable receivers appearing in the shops. However, with such enthusiasm for HDR, those timescales may be shorter this time around.

9. When you look back on your project in 20 years' time, what parts of it will you most remember?

AC: IBC2015 is clearly a turning point for Hybrid Log-Gamma. Broadcast engineers around the World are looking to see how they can deploy HDR services in a cost-effective manner, and with the alternative solutions that’s looking quite difficult. IBC provides us with an excellent opportunity to showcase the elegant simplicity of the Hybrid Log-Gamma solution. So I’m sure this year’s IBC will be a highlight.

One of the other highlights came just a few weeks ago, when a small group of us sat with a professional colourist and watched her grade Hybrid Log-Gamma HDR content for the show. Until that point all we’d seen were engineer’s colour grades, and whilst I don’t mean to criticise my immediate colleagues, we don’t have those craft skills and the results are like “chalk and cheese”.

TB: As always when you look back it is the people and the fun times you have had with them that you remember much more than the technology. We have been privileged to work with a great bunch of people from different organisations and across the world. Even when we haven’t agreed with them it has been a pleasure to work with them

Selected content from IBC2015

More is more: investigating attention distribution between the television and second screen applications - a case study with a synchronised second screen video game

R. Bernhaupt   R. Guenon   F. Manciet   A. Desnos

ruwido austria gmbh, Austria & IRIT, France

Abstract: Attention is a key concept for the design of second screen applications that are to be used while watching television (TV). One of the key design goals is to balance the user's attention between the second screen application and the TV content. To investigate the influence of interactivity on attention and overall perceived workload and user experience, we developed video games with varying degrees of interactivity allowing users to play while watching a TV show. The small games were synchronized with the TV show in both the temporal dimension (presenting games as the show progressed) and with the content (enabling users to play games similar to the storyline in the TV show). Results from a laboratory-based user study show that highly interactive video games that are interleaved with the content, draw up to seventy percent of attention on the game (tablet) and are judged with a perceived workload similar to doing real 'work' while watching TV. Nevertheless participants rated the games as fun to play with a high user experience.

Introduction

Second screen applications are becoming more and more popular and widespread due to the growing popularity of tablets and smart phones. These applications can considerably enrich the TV experience by providing more information or additional features, like background information about actors and movie plots, access to the movie's web site, or other interactive features like quizzes or voting.

Second screen applications typically allow the user to interact with the content they are consuming such as TV shows, movies or music. The main characteristic is that additional information is displayed on a portable device - usually the smart phone or tablet. The way that the content is enhanced (i.e. added value for the user) seems critical for the user acceptance. Work by Basapur et al. (2012) indicates that the experience is only positive when the additional media are in sync with the TV show and that it provides information that is judged to be relevant. The primary and secondary screens have to give the user the feeling of having a holistic (and synchronized) experience.

As Figure 1 presents, content on the tablet can be classified by degree of interactivity and by degree of synchronization. From the user's point of view, a high level of synchronization means having a holistic experience when using the TV screen together with a second screen application. An example is a game where you play along with the main character on the screen; the game being only available during a specified time or scene during the show. A low level of synchronization would describe an application that provides a TV experience only on the second screen without taking into account the content on the main TV screen. We refer to this category as TV stand-alone apps for second screens.

In terms of perceived interactivity, a high level of interactivity refers to a user interface design that allows the user to seamlessly interact with content on both screens (e.g. a jump-and-run game that is connected to the main screen content - and ideally would change the course of events in the movie or show). A low or limited level of interactivity would refer to a simple selection of content on the second screen (e.g. an electronic programme guide that could enable parts of the content to be displayed on the TV screen).

What is still unclear when designing for the second screen is what level of interactivity and what degree of synchronization lead to what level of attention. Is the TV becoming superfluous? And finally, what level of attention and overall perceived workload will be just too much and overwhelming for the user? Overall, what we would want to

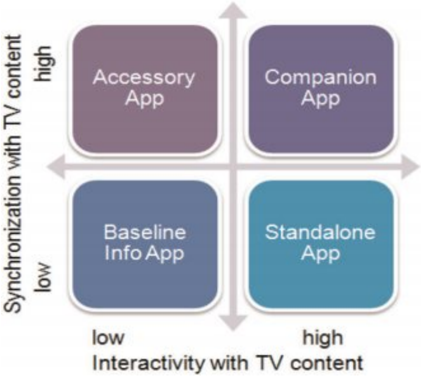


Figure 1 *Classification of second screen TV apps (from Bernhaupt et al., 2013).*

achieve is a positive user experience. But is this still the case when second screen applications are highly demanding?

The goal of this study was to understand how to balance the user's attention in a multi-screen environment by investigating the relationship of games with different degrees of interactivity and synchronization, measuring attention, perceived fun, difficulty of the game and overall perceived user experience in terms of hedonic and pragmatic quality. Our hypothesis was that watching a high-paced TV show while playing small games with varying levels of interactivity and high synchronisation with the content, would increase the perceived workload and therefore affect fun and user experience.

State of the art

While watching TV, people use a variety of other devices. During the past decade the devices used changed from standard household appliances and activities (e.g. ironing) (Bernhaupt et al., 2007) to more entertainment-oriented ones: Hess (2011) reports that people perform activities associated with the TV programme currently being watched, like searching the web to obtain information on the show, but also a variety of activities unrelated to the TV programme, including usage of social networks (e.g. Facebook, Twitter) or playing games.

When introducing second screen applications and performing activities on a second device, one key dimension for the design is how to control the user's attention between the two devices. Attention is a concept that refers to how we actively process specific information present in our environment. It implies that, of the multitude of possibilities to which the user can direct his/her attention, some are ignored in order to deal effectively with others (James, 1880).

Attention *per se*, is measured using behaviour-oriented measurements like the identification of patterns. Visual attention for example is measured with the D2 test (Brickenkamp, 1962). The visual scanning performance is evaluated by asking people to identify in a sequence of the letter "d" all the cases where the letter is accompanied by two small strokes. When measuring attention for interactive systems, researchers have been applying a multitude of methods ranging from standard tests to bio-physiological measurements. Currently eye-tracking is popular. Results from Hawkins et al. (1997) show that the average look at the television is about 7 seconds, with the median length of gaze being 2 seconds. When interacting with a second screen, Holmes et al. (2012) found that 63% of gaze time was on the TV, compared to 30% on the tablet and 7% off-screen. When using a second screen, gazes are becoming shorter than in the traditional settings, averaging under 2 seconds for gazes on the TV and just over one second on the second screen (Holmes, 2012). The main indicators from a user perspective are the overall hedonic quality of the system (aesthetics, identification and stimulation) as well as the emotional state of the user (e.g. fun).

For the design of TV user interfaces and second screen applications it is important to understand how visual attention can affect memory. As Holmes (2012, p.2) has stated: "For example, audio combined with text appears to overwhelm the separate channels, whereas combining audio with a related visual requires less time to process the information with increased performance." Following Paivio (1986) memory can be enhanced if visual and verbal information are paired and available in working memory. To understand how to enhance second screen applications it is thus necessary to evaluate the working memory e.g. via the task-load. A task load is a subjective value that indicates the perceived workload of tasks.

Playing along with the content: an experimental approach

Research question and method choice

The goal of this study was to investigate the effect of levels of interactivity for highly synchronized second screen games on attention and perceived fun and overall how synchronized games are perceived in terms of workload and user experience. Our hypothesis was that a high paced TV show along with mini-games with low to high levels of interactivity would be (even for the young generation) too high in workload to be enjoyable.

To investigate this research area we decided to perform a set of experiments starting with a small-scale user study, verifying our initial hypothesis on what the maximum of interactivity in terms of workload is for a young age group. The experimental oriented design of this study thus varied the degrees of interactivity within the game, and measured self-reported levels of attention and perceived workload using the NASA-TLX questionnaire.

To investigate the influence of the varying degrees of interactivity and high synchronisation we decided to perform an experiment combining a high-paced TV show and a range of small games (mini-games) with low to high degrees of interactivity. The level of attention needed for each activity (watching TV and playing the mini-game) was tuned in terms of activity in the gameplay to investigate how attention gets impacted and so how the overall workload of these games is perceived.

To evaluate attention we used participants' self-judgment as well as observation. The workload was measured using the NASA-TLX scale (Hart & Staveland, 1988). The NASA Task Load Index (NASA-TLX) is a subjective, multidimensional assessment tool that rates perceived workload. A set of different scales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration) are combined into one measure ranging from 0 to 100, the task load index.

As games that are not fun to play might affect the effort that people are putting in the game, and as systems that are not having the necessary positive user experience can limit overall engagement, we observed perceived fun and overall user experience. Fun was measured after each mini-game by self-evaluation (scale 1-10) and overall user experience was measured using the AttrakDiff questionnaire (Hassenzahl et al., 2004). The AttrakDiff is a questionnaire measuring pragmatic and hedonic quality (in terms of stimulation and identification as well as attractiveness) using a 7-scale semantic differential with word-pairs (e.g. ugly vs. attractive). Results are combined in two measures (pragmatic quality vs hedonic quality) and represented as a diagram showing the attractiveness of the product.

Prototype, participants and procedure

To understand the level of attention and perceived workload when playing along a TV show, we developed a set of three mini-games synchronized with a movie currently running on the TV screen. Based on the BBC Sherlock Season 1 Episode 1 TV show, a set of mini-games was developed to be played along. The mini-games were made available on the second screen app depending on the progress in the show (see Figure 2).

In the study, three mini-games were played. Game 1 (Clues) was a puzzle game where people had to find clues to solve the problem that was in line with the story (see Figure 3). Synchronization of the game and the content was very high (available time for the game was linked with the show progressing) and interactivity was low. Game 2 (Explore) was an exploration game where people had to find clues in a 3D representation of the streets that were at the same time used in the show, with a medium level of synchronization and interactivity. Game 3 (Jump-and-Run) was a game where the player had to run along with the main character on the screen to simulate catching a taxi. It had high synchronization and high level of interactivity.

Eight participants, five male, three female, aged 16 to 26 with a mean age of 22 years, took part in the study. The size of the eight households in which the participants lived ranged from one to four members. Participants reported that they played video games regularly on a multitude of devices. Seven participants reported that they had IPTV at home (covering all major operators in the country). Six

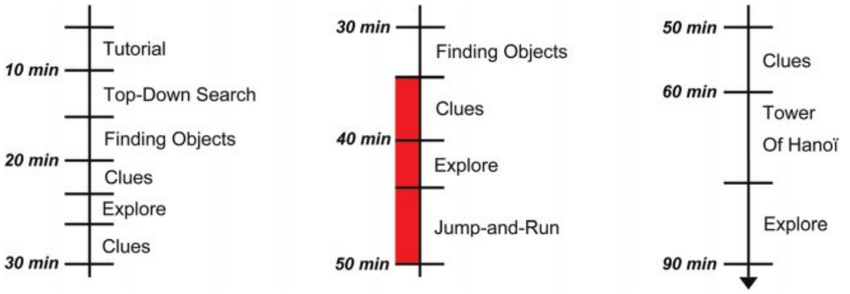


Figure 2 *Timeline of the Sherlock episode with different mini-games associated. In red, section of the episode retained for the prototype.*

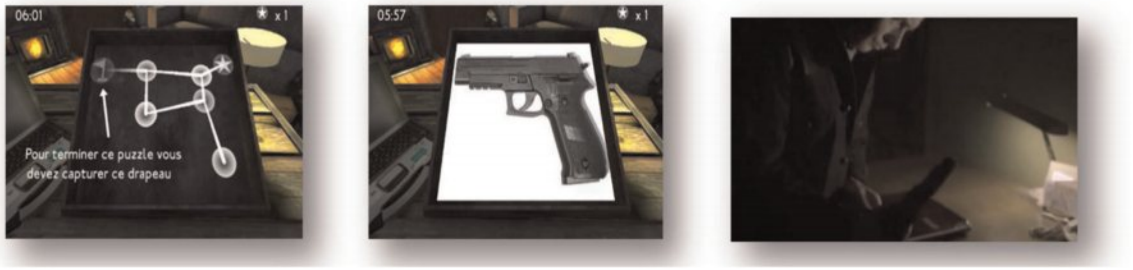


Figure 3 *Clues game: The player had to play Puzzle while main character in the TV show found the same clues.*

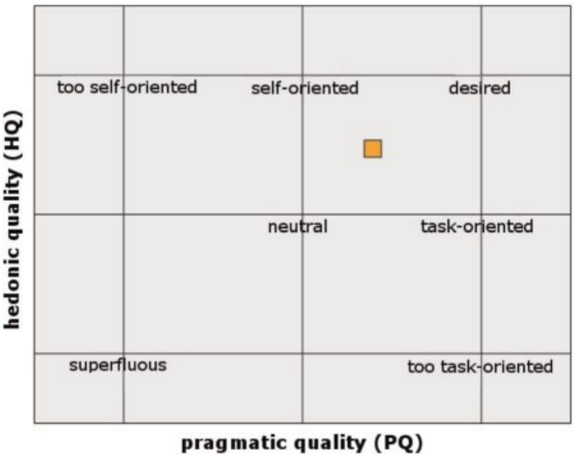


Figure 4 *AttrakDiff results for hedonic and pragmatic quality.*

participants reported watching TV (on a TV screen) at least several times per week, one reported watching once per week, one reported to watching once per month.

All participants used the Internet, mainly to check e-mails, play games and occasionally to do online shopping. While all participants owned a PC or laptop and a mobile phone, only four participants owned, or had access to, a tablet. None of the participants had used a second screen application while watching TV, before taking part in the study.

Participants were welcomed for the experiment and were seated in a low chair looking at a TV screen; a position that would be typical for watching TV on a couch. Participants signed the participant video allowance, consent form and answered some short questions on their TV watching and gaming behaviour. Participants then were handed over the tablet with the games application already running, and the Sherlock Episode was played on the TV screen. To minimize duration participants were asked to play along with three segments of the 90 min show (red in Figure 2).

After each of the three mini-games, participants had to judge their perceived level of fun, degree of attention to the game, interconnectedness between TV content and game and overall difficulty of the game. At the end of the episode, participants filled out the NASA-TLX and the AttrakDiff questionnaires and answered a set of interview questions related to their experiences with the system. Events in the game (on the tablet) were logged, and participants were video-recorded with a camera. We recorded the user from the front to focus on both the user's face and hands interacting with the tablet.

For the analysis we had two researchers investigating the videos and coding on an agreement basis. All material used, including video and games, were in French, results from interviews were translated by two researchers independently.

Results

All participants playing the game started interacting with the game within the first 10 seconds after the game started, on average after 4.9 seconds (SD = 1.9). Each of the three games was played by the eight participants once. From the 24 game sessions played, 22 were played successfully (reaching the game goal within the given time), one player ran out of time while playing the game, while one result was missed due to technical logging problems. While playing the three mini-games (Clue, Exploration and Jump-and-Run) none of the participants reported any problems or difficulties or asked any questions.

Table 1 shows the results for the three different mini-games (G1, G2 and G3) for overall attention and users' self-judged evaluation of fun and difficulty of the game. Results are contrasted with observer-judgements of the attention ratio between TV and tablet.

What can be seen in Table 1 is that the design of the mini-games fitted the purpose: G1 was perceived low in difficulty (low interactivity), and the synchronization with the main TV content was high. Attention to the TV screen was still high with 46%. For G2, the exploration game, with medium synchronization and interactivity, attention to the TV was rated 37.5%, while perceived difficulty was rated medium, and fun was rated rather high (7.5 on average on a scale from1 to 10). For G3, the Jump-and-Run game, synchronization of the content and interactivity were high, leading to low attention for the TV screen, high scorings for fun and a rather low difficulty rating of 3.7. A one-way between subjects ANOVA was conducted to compare the effect of degree of interactivity in the game. There was a significant effect of interactivity on fun at the $p < .01$ level for the three games ($F(2.21) = 7.76$, $p < 0.01$) and for attention ($F(2.21) = 5.29$, $p < 0.05$). Difficulty was not significant, but there was a clear trend visible ($F(2.21) = 10.71$, $p = 0,051$).

Table 1 - Results of observed attention on TV, perceived fun and difficulty of the games.

G1 (Clues)			G2 (Explore)			G3 (Jump-and-Run)		
Attention on TV (%)	Fun (1-10)	Difficulty (1-10)	Attention on TV	Fun	Diff.	Attention on TV	Fun	Diff.
46%	6.1	2.5	37.5%	7.5	4.8	30%	7.9	3.7

Table 2 - Showing results for the six scales of the NASA-TLX and the overall workload score (on a scale from 0 to 100).

Mental Demand	Physical Demand	Temporal Demand	Performance	Efforts	Frustration	NASA-TLX (workload)
62.8	31.8	39.4	56.2	47.5	38.7	59.12
min: 35 max: 75	min: 5 max: 75	min: 10 max: 75	min: 10 max: 85	min: 30 max: 65	min: 5 max: 65	SD = 3.29

The NASA-TLX questionnaire measures perceived workload on a scale from 0 to 100, with results around 60 indicating a medium workload, or an activity similar to a office work (Cinaz et al., 2013). It replicates results from users watching TV and performing interleaved work-tasks indicating workloads around 60 (Du Toit, 2013).

In terms of user experience, measured with the AttrakDiff questionnaire (Hassenzahl, 2004), the second screen application was rated above average regarding the pragmatic and hedonic quality on the 7 point scale, while failing to achieve a rating to be really desired. This can be interpreted as while users perceived the user experience in terms of hedonic and pragmatic quality as acceptable, there is still a lot of room for improvement to reach an overall experience that is really desired by the user.

The results show that high attention games are judged positively by users in terms of fun and perceived workload is about the same degree as people would have when performing an interleaved work-task. What we found is that the more the gaming action is synchronized with the movie's storyline, the easier the player can follow the two media streams simultaneously, which is in line with the psychological theory on two-cue coding (Paivio, 1986). On the contrary, highly synchronized games lead to a shift of attention away from the TV to the tablet. What we found in this particular case study (with a rather high-paced TV show, high degree of synchronization and high degree of interactivity) is that the attention of the player is on the tablet and not any longer on the TV show, with a worst case of only 30% of attention on the TV.

Summary and discussion

How much attention to the second screen is too much attention? Based on results from this experimental study with young participants (age 16 to 26 years), we found that second screen games that are interleaved with the programme draw the majority of attention to the second screen (up to 70%). But they are still judged as fun to play and to provide a good overall user experience. What has to be critically discussed is if such a finding can be generalized to other user groups and applications. As indicated by the workload measure, playing an interleaved game is perceived as the same level of workload as performing actual work. This might indicate that for short periods of gameplay a high level of workload might be acceptable, but that it is less likely to be accepted for longer periods. For younger age groups playing such game might be interpreted as a challenge and they really want to get immersed in the game, but for other applications like additional information or social connectedness, this level of workload will very likely not be accepted.

For the design of second screen applications it is advisable to interleave the content as much as possible, with clear indications where the user should look at. This way attention can be controlled and attention for the screen can be better balanced. To limit the workload it can be helpful to interleave the activities on the two screens so the activity of watching and playing feels like doing one task, instead of two separate ones. Separate tasks have been shown to have load ratings up to 80 (Du Toit, 2013), which might not be advisable for an entertainment application. To summarize, contrary to lots of design advice for second screen activities, the more activity is expended on the second screen (in line with the content) the better the overall user experience.

References

[1] BASAPUR S., MANDALIA H., CHAYSINH S., LEE Y., VENKITARAMAN N., METCALF C.: 'FANFEEDS: evaluation of socially generated information feed on second screen as a TV show companion', *Proceedings of the 10th European conference on Interactive TV and Video*, July 2012, pp. 87–96

[2] BERNHAUPT R., OBRIST M., WEISS A., BECK E., TSCHELIGI M.: 'Trends in the living room and beyond', *Interactive TV: a Shared Experience*, 2007, pp. 146–155

[3] BERNHAUPT R., PIRKER M., GATELLIER B.: 'Identification of user experience and usability dimensions for second screen applications: results from an expert evaluation using generic task models', *Proceedings of 2013 International Broadcasting Convention*, 2013

[4] CINAZ B., ARNRICH B., LA MARCA R., TRÖSTER G.: 'Monitoring of mental workload levels during an everyday life office-work scenario', *Personal and Ubiquitous Computing*, 2013, Vol. 17, No. 2, pp. 229–239

[5] DU TOIT H.: 'Working while watching TV, is it really work?: the impact of media multitasking on stress and performance', 2013

[6] HART S.G., STAVELAND L.E.: 'Development of NASA-TLX (task load index): results of empirical and theoretical research', *Advances in Psychology*, 1988, No. 52, pp. 139–183

[7] HASSENZAHL M.: 'The interplay of beauty, goodness, and usability in interactive products', *Human-Computer Interaction*, 2004, Vol. 19, No. 4, pp. 319–349

[8] HAWKINS R., PINGREE S., BRUCE L., TAPPER J.: 'Strategy and style in attention to television', *Journal of Broadcasting and Electronic Media*, 1997, No. 41, pp. 245–264

[9] HESS J., LEY B., OGONOWSKI C., WAN L., WULF V.: 'Jumping between devices and services: towards an integrated concept for social TV', *Proceedings of the 9th International Interactive Conference on Interactive Television*, 2011, pp. 11–20

[10] HOLMES M.E., JOSEPHSON S., CARNEY R.E.: 'Visual attention to television programs with a second-screen application', *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2012, pp. 397–400

[11] JAMES W.: 'The principles of psychology, Vol 1' (Dover Publications, 1950, reprint from the Original Article in 1880)

[12] PAIVIO A.: 'Mental representation: a dual coding approach' (Oxford University Press)

Interview - Raphaël Guénon and François Manciet

1. Tell us a bit about yourself and what you do

RG: I graduated with a Master's degree in Computer Science (Image Synthesis) from University Toulouse III - Paul Sabatier in 2014.

In my role as research engineer my main goal is to develop new interactive TV user interfaces and second screen applications based on development tools and techniques from the game industry. I am currently working at IRIT in Toulouse, France.

I am developing the leaf UI, a TV UI aiming to synchronize the input device with the interface. In the next few years I will be working on several projects, trying to investigate various fields such as cross-device interface, attention distribution and new interactive techniques.

FM: I graduated with a Masters degree in Human-Computer Interaction from University Toulouse III - Paul Sabatier in 2014.

I am currently a first year PhD student. My main goal of my thesis is to investigate the current behaviours in the living room related to entertainment and what types of tasks and activities can be changed to enhance the user experience, applying automation and task migration.

One of my research topics is to design and assess in a user-centered development manner new possible interaction techniques that help users when crossing-over using devices of the living room (TV, smartphones, tablets). I am currently doing my thesis at IRIT in Toulouse, France.

2. What recreational activities do you enjoy outside of work?

RG: Playing games - I currently play 'The Talos Principle' and 'L.A. Noire'. I like to play basketball but to really relax my favourite activity is drawing and sketching.

FM: My way to relax is to run and also to play basketball - and of course I play video games.

3. Do you use a second screen while watching TV? And how would you say this enhances your enjoyment of a programme?

RG: Honestly - no. For me, what you call the second screen is the major screen for most of my activities, including video consumption.

FM: As I am still a student, my main second screen is my smart phone, where I continuously look up information related to the movies and TV shows that I watch together with my girlfriend. It really helps me to find out which actor she is again "swooning" over :-)

4. Tell us briefly what your work revealed about playing a demanding video game while watching TV

RG: The central message is that synchronization is the key to successful second screen application. If the content on the second screen is not perfectly in line and synchronized it is ways less engaging for the consumer.

5. Many will be surprised by the conclusions of your experiments, were you?

FM: A bit. We initially thought that the easier game would be the one preferred by the users, but in the end it was the jump-and-run game, that is more complicated and more demanding, that got a better rating. But of course this cannot be generalized as a result: it does not mean to build in a runner game for everyone, but to build an experience that is really immersive.

6. What kind of TV programmes are best and least suited to second screen enhancement?

RG: It is not so much about the program, it is about the targeted population of gamers: not everyone likes the same games, so the decision is more, what type of gamer do you want to support with your game alongside the programme, and maybe even, can you make several games that fit for different groups of gamers?

7. Do you think that the artistic value of the main programme will be degraded if accompanied by a second screen application?

FM: Goal is to have the same artistic value on the second screen, and this is where lots of second screen applications today are failing. But reproducing the same artistic value on the second screen will add up and provide a better user experience.

8. Do you avoid audio in a second screen application because the viewer must rely on main programme audio to follow what is going on?

RG: Yes, exactly. This allowed us to connect the two screens, as you use the sound of the movie as sound in the game.

9. Your experiments used people aged from 16-26 years, do you think that older viewers will embrace second screen use and enjoy high-demand applications?

FM: Older viewers typically are interested in other types of games, and a runner game might not be very attractive for them. Nevertheless I assume that high-demand applications that provide an immersive experience can be designed for all age groups. But - it is not one game or application - it will be several ones.

10. Broadcasters have been slow to introduce second screen enhancement, why do you think this is?

RG: I spent a lot of time and effort to design, develop and evaluate the small games that we used in the experiment. It was a huge effort to make these games good enough in terms of synchronization. I assume broadcasters might fear the investment they have to make, to reach such a high level of synchronization for the big amount of material they might want to support with second screen games.

11. UHDTV is more of an immersive and involving experience; will this mean that viewers will spend less time with their second screens?

FM: UHDTV in terms of technology will help to provide even more immersive experiences, but I fear it will take some time that UHDTV will reach the end consumers.

12. TV watching has been a shared and social activity,

won't second screen gaming make it more solitary?

RG: Yes and no. What we did not present in the paper is that for example the clues game is designed in a way that it can be played by two people on the same tablet, and that especially the jump-and-run game is a good candidate to be played on two tablets in a kind of competition. In our overall concept we carefully chose the way to design the game: Playing the game does not negatively impact the TV watching experience of someone not using a second screen, and the second screen content is only partly available so communication with other people watching TV with you is enabled.

13. Should the main and second screen become a single hand-held screen? Wouldn't that make life easier?

FM: Examples of how one single hand-held screen can be used to have content displayed that integrates the game do already exist, but overall they are not very successful. I personally think two screens just allow more information - and this is what the younger generation is currently enjoying and looking for.

14. Of all the emerging ways of enjoying visual media: head-mounted panoramic, UHD, interactive second screen, phone on the move, etc which do you find most exciting?

RG: I would love to play around with the ‘Oculus Rift’ and explore its possibilities for video games and TV. Especially when you create video games it is about building a world, and it just sounds fascinating to enable users to visit and explore such a virtual world.

FM: For me most exciting are strategies like the Internet of Things that allow the seamless access to content anytime, anywhere and on any device. I worked on enabling people to take their content to any type of device simply touching it with the remote control. This way people have an easy way to access and distribute their content between different devices (and operating systems), by supporting providers to do the rights management via the bi-directional remote that can handle access. It is fascinating to enable users to have access to all these connected devices.

15. Will you both continue to work in the media industry and where would you like your careers to take you?

RG: Since finishing my Master I work for ruwido developing future and futuristic concepts supporting entertainment and media consumption. On average people employed at ruwido work there around 15 years, so I assume I will be able to come up with a series of interesting ideas on how people are entertained in the future.

FM: To understand how to better support media consumption in the living room, I think it is important to understand concepts like automation, migration of tasks and connectivity. This is why I started a PhD in Computer Science. On the long term my hope is to work in different places all over the world.

Selected content from IBC2015

HEVC/H.265 codec system and transmission experiments aimed at 8K broadcasting

Y. Sugito[1] *K. Iguchi*[1] *A. Ichigaya*[1] *K. Chida*[1] *S. Sakaida*[1]
H. Sakate[2] *Y. Matsuda*[2] *Y. Kawahata*[2] *N. Motoyama*[2]

[1] *NHK, Japan*
[2] *Mitsubishi Electric Corporation, Japan*

Abstract: This paper introduces the world's first video and audio codec system that complies with 8K broadcasting standards and describes transmission experiments via a broadcasting satellite using this system.

8K Super Hi-Vision (8K) is a broadcasting system capable of highly realistic 8K Ultra High Definition Television (UHDTV) video and 22.2 multichannel audio. In Japan, domestic standards for 8K broadcasting were formulated in 2014 and 8K test broadcasting will begin in 2016. We have developed an 8K High Efficiency Video Coding (HEVC)/ H.265 codec system that complies with the domestic standards.

In this paper, we first explain the features and a roadmap for 8K broadcasting. We then introduce the features of the 8K HEVC/H.265 codec system developed. Finally, we describe transmission experiments using a satellite system that is equivalent to the one that will be used in test broadcasting. The results allowed us to confirm that the developed system provides high-quality transmission at the expected bit rate during test broadcasting.

Introduction

An 8K Super Hi-Vision (8K) broadcasting system capable of highly realistic 8K Ultra High Definition Television (UHDTV) video and 22.2 multichannel (22.2 ch) audio is currently under development. In Japan, domestic standards for 8K broadcasting were formulated in 2014 and 8K test broadcasting using a broadcasting satellite will begin in 2016. The standards prescribe video coding, audio coding, multiplexing and transmission schemes, and other such procedures. Compression of video data to a transmittable bit rate while maintaining high quality is a key issue that must be addressed in order to realize 8K broadcasting.

A new video coding scheme, referred to as High Efficiency Video Coding (HEVC)/H.265 (1), was standardized in 2013. HEVC supports 8K video formats and achieves approximately twice the compression level of the existing Advanced Video Coding (AVC)/H.264 scheme. Its coding performance shows a particularly significant improvement relative to the previous schemes for high-resolution video such as 8K. On the other hand, the computational cost for HEVC encoding and decoding is quite high and is reported to be more than twice that of AVC. This makes realization of real-time HEVC processing challenging.

We have developed the world's first 8K HEVC/H.265 real-time codec (encoder and decoder) system that complies with the domestic standards. The system allows 22.2 ch audio coding and video/audio multiplexing, with these functions integrated into the video encoder and decoder. The system was tested by transmission experiments using a broadcasting satellite in 2015.

In this paper, we first explain the features and a roadmap for 8K broadcasting. We then introduce the features of the 8K HEVC/H.265 codec system. Finally, we describe transmission experiments using a satellite system that is equivalent to the system that will be used for the test broadcasts.

8K broadcasting

In this section, we explain the features and a roadmap for 8K broadcasting.

8K Super Hi-Vision

8K is a TV broadcasting system designed to deliver highly realistic 8K video and 22.2 ch audio. Table 1 shows the parameters for 8K and a current 2K digital broadcasting system using a broadcasting satellite (BS) in Japan. The 8K provides much higher fidelity than the existing broadcasting system.

The 8K video format is internationally standardized in Recommendation (Rec.) International Telecommunication Union Radiocommunications Sector (ITU-R) BT. 2020 (2). Since the number of pixels in the horizontal direction is 7,680 pixels, it is called "8K." Its characteristics are higher spatial resolution, higher frame rate, higher bit

Table 1 Parameters for 8K Super Hi-Vision and current BS digital broadcasting in Japan

		8K Super Hi-Vision	2K BS digital broadcasting
Video	spatial resolution (pixels in H × V)	7,680 × 4,320	1,920 × 1,080
	aspect ratio	16:9	16:9
	scan mode	Progressive	Interlaced
	frame rate (Hz)	120, 120/1.001, 60, 60/1.001	60/1.001
	bit depth (bit)	10, 12	8
	color gamut	Rec. 2020	Rec. 709
Audio	number of channels	22.2	maximum 5.1
	sampling rate (kHz)	48, 96 (optional)	32, 44.1, 48
	quantization bit (bit)	16, 20, 24	16 or more

depth, and wider color gamut than those of the existing broadcasting video formats. The uncompressed bit rate for the 8K/12 bit/60 Hz format is approximately 72 Gbps.

The 22.2 ch audio format is standardized in Society of Motion Picture & Television Engineers (SMPTE) 2036-2-2008 (3). It is characterized by a larger number of channels, increased sampling rate, and larger quantization bit. The uncompressed bit rate for the 22.2 ch/48 kHz/24 bit format is approximately 25 Mbps.

Because of the enormous uncompressed bit rate of 8K video, compression of the video to a transmittable bit rate while maintaining high quality is a key issue for 8K broadcasting.

8K broadcasting roadmap

In Japan, a committee of the Ministry of Internal Affairs and Communications (MIC) presented a roadmap for 8K broadcasting in 2014. According to the roadmap, 8K test broadcasting will begin in 2016 and 8K broadcasting is planned to begin by 2018. Broadcasting satellite system will be used for transmission in both broadcasting types. Moreover, 8K broadcasting is expected to become common in 2020, the year of the Olympic and Paralympic Games in Tokyo, Japan.

8K HEVC/H.265 codec system

We have developed the world's first 8K HEVC/H.265 codec (encoder and decoder) system that complies with the 8K broadcasting standards.

Figure 1 shows a photograph of the system. The system is composed of an encoder to compress video and audio data, a decoder to decompress encoded video and audio data (upper silver devices in the racks), and interface converters for encoder input and decoder output. The 8K HEVC encoder was developed in 2013 as reported by Sugito et al. (4), and

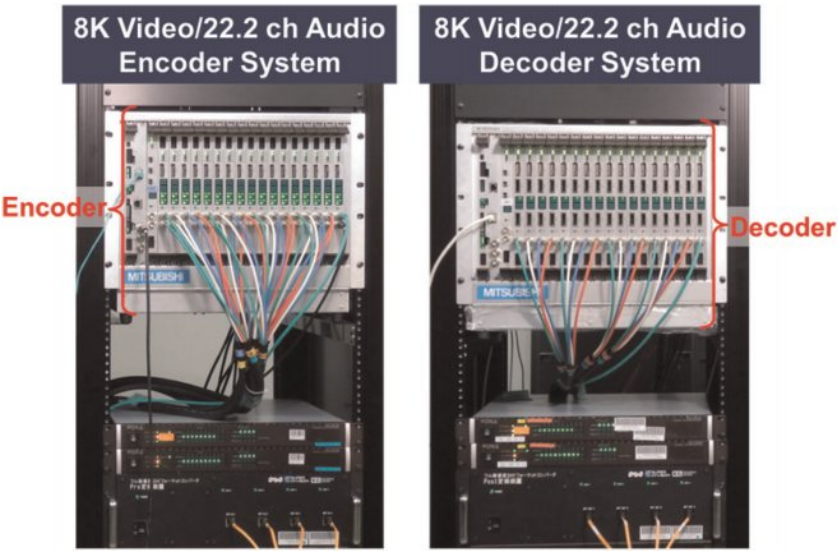

8K Video/22.2 ch Audio Encoder System
8K Video/22.2 ch Audio Decoder System
Encoder
Decoder
MITSUBISHI

Figure 1 *8K HEVC/H.265 codec system*

Table 2 8K HEVC/H.265 codec system specifications

Video	coding scheme	MPEG-H HEVC/H.265 Main 10 profile@ Level 6.1
	spatial resolution and frame rate	7,680 × 4,320/59.94 P
	chroma format and bit depth	4:2:0/10 bit
	input/output interface	3G-SDI × 17
Audio	coding scheme	MPEG-4 AAC Low Complexity (LC)
	number of input/output channels	22.2 ch
	sampling rate and quantization bit	48 kHz/24 bit
	input/output interface	MADI (AES10)
Multiplexing	multiplexing scheme	MPEG-H MMT
	input/output interface	RJ-45 × 1

the 8K HEVC decoder has been newly developed in 2015. Table 2 shows the system specifications. In this section, we introduce the features of the new system.

Real-time 8K HEVC video encoding and decoding

The codec is capable of real-time 8K HEVC video encoding and decoding, which require an extremely large number of calculations. Figure 2 is a diagram of the 8K HEVC encoder.

To enable real-time processing, each video frame is spatially divided into 17 horizontal strips, and these strips are encoded in parallel. The encoder consists of 17 encoding boards. Each encoding board processes a single strip, and adjacent boards share the motion information needed for encoding. This method of partitioning was chosen based on a number of factors, including the transmission capacity of shared motion information between encoding boards, the pixel count for the 3G-SDI standard, and the convenience of a wider horizontal motion search range.

The codec system adopts HEVC/H.265 (1) as the video coding scheme. HEVC is the latest video coding scheme standardized in 2013; its coding performance is significantly improved compared to previous schemes, especially for high-resolution videos such as 8K. This is because HEVC has many more coding modes compared to the previous video coding schemes. On the other hand, the availability of many coding modes can lead to difficulties in selecting the most appropriate mode, and this is one of the current problems for real-time processing using HEVC. To address this, the encoder automatically determines the

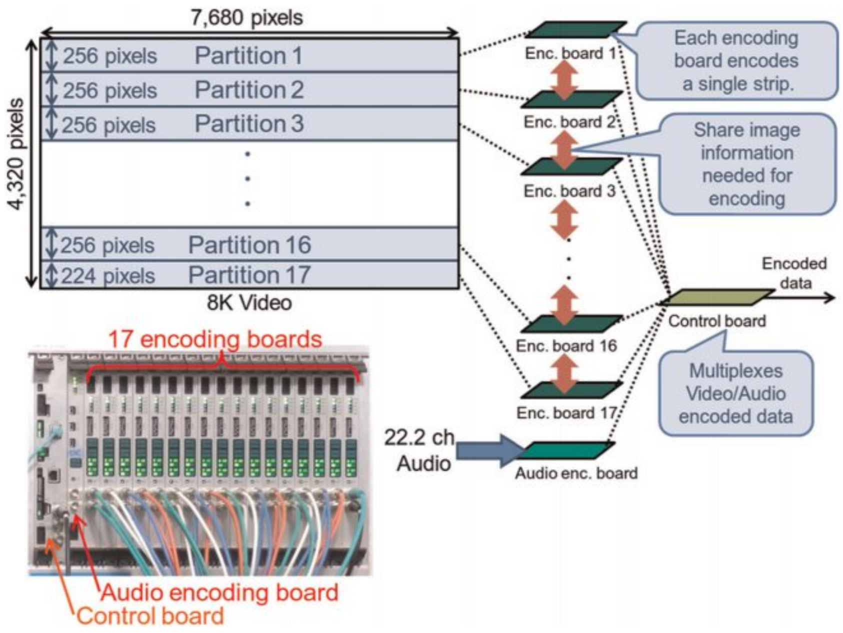


Figure 2 *A diagram of the 8K HEVC encoder*

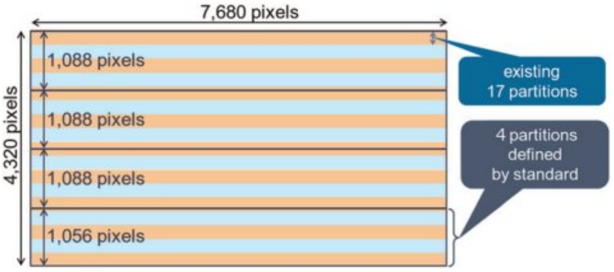


Figure 3 *Encoder modification based on domestic standard*

coding parameters depending on the complexity of the original image, coding result, and parameter setting.

The newly developed decoder is implemented following a similar design. The decoder is composed of 17 decoding boards, and each decoding board processes a spatial partition of the video frame in parallel.

Audio codec and multiplexing capability

The codec system includes a 22.2 ch audio encoder/decoder using Moving Picture Experts Group (MPEG)-4 Advanced Audio Coding (AAC) (5) and a multiplexing/demultiplexing capability using MPEG-H MPEG Media Transport (MMT) (6) to combine and transmit compressed video and audio data.

As shown in Figure 2, the audio encoder is implemented by one board and integrated into the 8K HEVC encoder. The audio decoder is made in the same manner. The multiplexing capability is implemented as a function of the control board in the encoder. The board combines elementary stream (ES) from video and audio encoding boards and outputs encoded data in the MMT format. Demultiplexing capability is contained in the decoder control board that interprets the data in the MMT format and distributes ES for video and audio decoding boards.

Compliant with domestic standards for 8K broadcasting

This is the video and audio codec system that complies with the 8K broadcasting standards. In Japan, the domestic standards for 8K broadcasting were formulated by the Association of Radio Industries and Businesses (ARIB) in 2014. The standards prescribe video coding, audio coding, multiplexing and transmission schemes, and other such procedures. The video coding, audio coding, and multiplexing schemes of the system comply with the domestic standard ARIB STD-B32 ver.3.1 revised in December 2014.

The standard governs the 8K video encoding method, and we therefore modified the 8K HEVC encoder to conform to the standard. Figure 3 shows the encoder modification based on the standard.

For 8K video encoding, the standard mandates four horizontal spatial divisions; three of these consists of 1,088 pixels in the vertical direction and the fourth one consists of 1,056 pixels in the vertical direction. Since the standard allows subdivision of the required four partitions, we modified the encoder to subdivide the four partitions by the existing 17 partitions.

Transmission experiments

We conducted the world's first transmission experiments via a broadcasting satellite using the developed 8K HEVC/H.265 codec system. The experiments were exhibited in OPEN HOUSE at NHK Science & Technology Research Laboratory (STRL) in May 2015 (7). In this section, we describe the 8K experimental broadcasting system and the transmission experiments.

8K experimental broadcasting system

Figure 4 shows a diagram of the 8K experimental broadcasting system.

In the system, uncompressed video and audio is provided from a signal source. More specifically, the signal is generated from recorders or cameras and microphones. An encoder then compresses the video and audio to a transmittable bit rate and outputs the compressed data in the MMT format. Table 3 shows the video and audio compressed bit rate settings in ES and the compression ratio. Although the video and audio bit rate for 8K broadcasting is still under consideration, we chose the bit rate by taking the bit rate for data broadcasting, transmission header, and the transmission capacity of the satellite into consideration. Uncompressed video and audio formats in the system are 8K/59.94 Hz/12 bit and 22.2 ch/48 kHz/24 bit, respectively. The compression ratio for the video is very large because of the HEVC scheme.

Next, the compressed and the subtitle data are combined by an MMT Multiplexer (Mux) and are encrypted by a

Table 3 Compressed bit rate settings in the experiments

	Compressed bit rate (ES)	approx. uncompressed bit rate	approx. compression ratio
Video	85 Mbps	72 Gbps	840
Audio	1.4 Mbps	25 Mbps	20

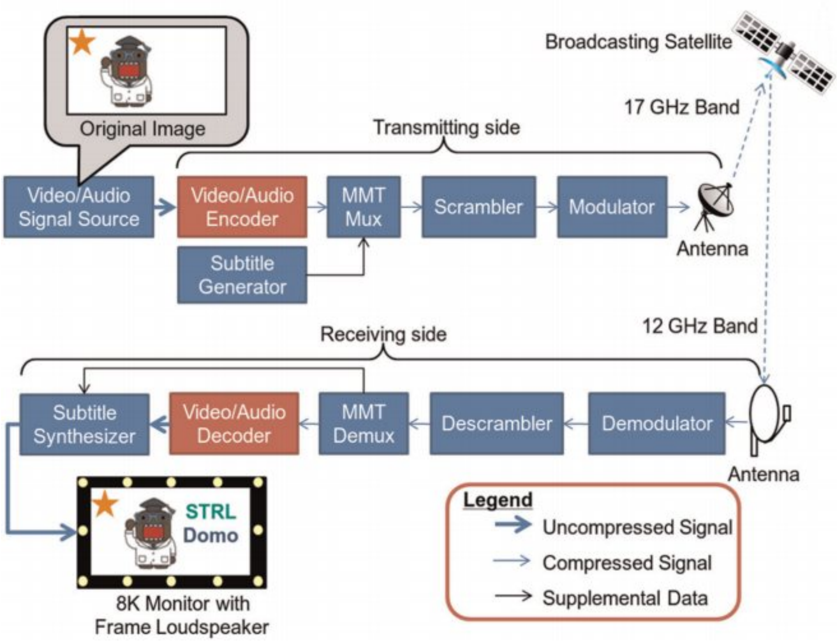


Figure 4 *A diagram of the 8K experimental broadcasting system*

scrambler. To adapt to the transmission path, the amount of the encrypted data is restricted to not more than 100 Mbps. Finally, the data is transmitted via a broadcasting satellite through a modulator and an antenna. The above devices are on the transmitting side and generally exist in a broadcasting station.

In TV transmission, the data for video dominates the transmission. As shown in Table 1, the uncompressed bit rate for 8K video is 32 times larger (16 times larger in spatial resolution and twice larger in temporal resolution) than that of the current BS broadcasting. In the system, 8K transmission is realized by adapting the HEVC scheme and transmission scheme. The compression performance of HEVC is approximately 4 times larger than that of the MPEG-2 video coding scheme used in the current 2K BS broadcasting. In the experiments, the broadcasting satellite is the same as the one currently used in the 2K BS broadcasting; however, because of the improvement in the transmission scheme, it can transmit twice as much data (up to approximately 100 Mbps) as the existing system studied by Suzuki et al. (8). The transmitting scheme is defined by the domestic standard, ARIB STD-B44 ver.2.0, in 2014. The satellite used in the experiments is planned to be used in 8K test broadcasting in 2016.

As shown in Figure 4, the transmitted data from the satellite is received by a demodulator through an antenna and decrypted by a descrambler. An MMT demux (demultiplexer) separates the subtitle data and the compressed data for video and audio. Video and audio are then expanded into the uncompressed data formats by a decoder, and a subtitle is synthesized on the video. These are on the receiving side and will be installed in consumers' homes by 2020.

Experimental results

We conducted transmission experiments for the 8K broadcasting system.

The results showed that 8K video and 22.2 ch audio are properly encoded, transmitted, and decoded. We confirmed that the developed system provides high-quality transmission at the expected bit rate during test broadcasting. We measured the delay time between the original video input and the decoded video display and found that it was approximately 3.5 s.

Conclusions

We have developed the 8K HEVC/H.265 codec system that complies with 8K broadcasting standards. We conducted transmission experiments with the system by using a broadcasting satellite. The results showed that the system is capable of 8K broadcasting.

References

[1] ISO/IEC 23008-2, 2013. High Efficiency Coding and Media Delivery in Heterogeneous Environments – Part 2: High Efficiency Video Coding | Recommendation ITU-T H.265, 2013. High Efficiency Video Coding

[2] Recommendation ITU-R BT.2020-1, 2014. Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange

[3] SMPTE 2036-2-2008, 2008. Ultra High Definition Television – Audio Characteristics and Audio Channel Mapping for Program Production

[4] SUGITO Y., IGUCHI K., ICHIGAYA A., CHIDA K., SAKAIDA S., SHISHIKUI Y., SAKATE H., ITSUI T., MOTOYAMA N., SEKIGUCHI S.: ‘Development of the super hi-vision HEVC/H.265 realtime encoder’, *SMPTE Conf. Proc.*, 2013, Vol. 2013, No. 10, ppp. 1–16

[5] ISO/IEC 14496-3:2009/Amd 4, 2013. Coding of Audio-visual Objects – Part 3: Audio

[6] ISO/IEC 23008-1, 2014. High Efficiency Coding And Media Delivery In Heterogeneous Environments – Part 1: MPEG Media Transport (MMT)

[7] http://www.nhk.or.jp/strl/open2015/en/tenji_1.html, 2015. 8K Satellite Broadcasting Experiment

[8] SUZUKI Y., TSUCHIDA K., MATSUSAKI Y., HASHIMOTO A., TANAKA S., IKEDA T., OKUMURA N.: ‘Transmission system for 8K super hi-vision satellite broadcasting’, *IBC2014 Conference*, 2014, p. 8.4

Selected content from IBC2015

Improving content interoperability with the DASH Content Protection Exchange Format standard

L. Piron[1] *K. Hughes*[2] *T. Inskip*[3]

[1]*Nagravision, Switzerland*
[2]*Microsoft, USA*
[3]*Google, USA*

Abstract: Content Protection is one of the key success factors in the deployment of an OTT TV system. To enable various sustainable business models, Service Providers need to securely and efficiently implement the interoperability of DRM-protected content across multiple devices.

Secured software client integration is needed at the device level for retrieving keys and decrypting content in a controlled environment. There are many different implementation frameworks for executing this function, going from pure software to hardware-based solutions. Secured integration is also needed at the streaming platform head-end, so that the DRM system core elements such as content keys, needed for encrypting content and also for generating content usage licenses, can be used at the right place and at the right moment in the head-end system.

One key aspect of the DRM ecosystem is that it encompasses multiple DRM vendors with specific implementations across a broad and growing pool of devices. A service provider thus needs to ensure that the components in its head-end, often coming from multiple vendors, can also support multiple DRM vendors to ensure the interoperability of content across multiple devices.

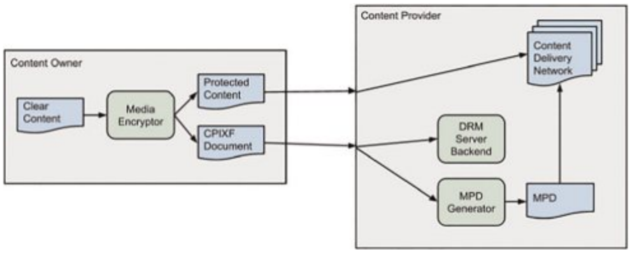
To help address this challenge, DASH-IF has defined a set of supported use cases and a secure container for exchanging content keys between DRM systems and head-end components such as encoders and CMS systems. This allows for the seamless exchange of content keys between all components in the head-end. Content Security remains in the hand of the operator, as such interfaces and key exchanges are secured. Leveraging the DASH Common Encryption standard (CENC), the same piece of content is encrypted once and used on many different devices with the appropriate key.

Introduction

The MPEG-DASH ecosystem is growing quickly, and a significant portion of the content being prepared and delivered is protected content. DASH profiles that have been deployed use MPEG Common Encryption for content protection in order to allow a single encoded and encrypted Representation to be played on multiple device types using multiple Digital Rights Management (DRM) systems. Common Encryption specifies standard content protection information in ISO Media Representations and DASH manifests such as key identifiers, DRM system identifiers, etc. that can be shared throughout the DASH ecosystem.

Preparation of protected media content for delivery may involve multiple entities and processing steps. For example, a content owner may encrypt some premium content and deliver it to multiple content providers, which in turn may generate their own DASH Media Presentation Descriptions (MPDs), and make the media decryption keys available to end users via their DRM server(s). Without a common interchange format for the copy protection information, each content owner, provider, and/or DRM system might specify their own, non-interoperable means of importing and exporting copy protection related data. The Copy Protection Information Exchange Format (CPIXF) specification aims to provide interoperability for these functions by standardizing the way in which entities and media processors exchange content keys and associated copy protection metadata.

The following diagram illustrates the example described above.



The Copy Protection Information Format specification fulfills the following requirements:

- Interoperability. This is the main goal of the specification; to allow the exchange of copy protection related data using a well-defined, specified and public mechanism.

- Flexibility. The CPIXF documents may be used in simple one-to-one exchanges, or in more complex workflows.

- Security. All security-sensitive information in the exchange (e.g. content keys) is encrypted in a manner such that only the intended recipient can decrypt it.

Content preparation workflows

Content keys and DRM signalization need to be created and exchanged between some system entities when preparing content. The flows of information are of very different nature depending on where content keys are created and also depending on the type of content that can be either on-demand or live for example.

The following gives a general overview of the context in which content protection information made of keys and DRMs signalization needs to be exchanged between entities in the backend.

Figure 1 shows logical entities that may send or receive DRM information such as content keys, asset identifiers, licenses, and license acquisition information. A physical entity may combine multiple logical roles, and the point of origin for information, such as content keys and asset identifiers, can differ; so various information flows are possible. This is an informative example of how the roles are distributed to facilitate the description of workflow and use cases. Alternative roles and functions can be applied to create conformant content. The different roles are:

Content Provider – A publisher who provides the rights and rules for delivering protected media, also possibly source media (mezzanine format, for transcoding), asset identifiers, key identifiers (KID), key values, encoding instructions, and content description metadata.

Encoder – A service provider who encodes in DASH format with specified media format, number of streams, range of bitrates and resolutions, seamless switching constraints, etc., possibly determined by the publisher. Each encoded track needs to be associated with a key identifier, a Representation element in an MPD, a possible 'pssh' box in the file header, and a DRM license separately downloaded.

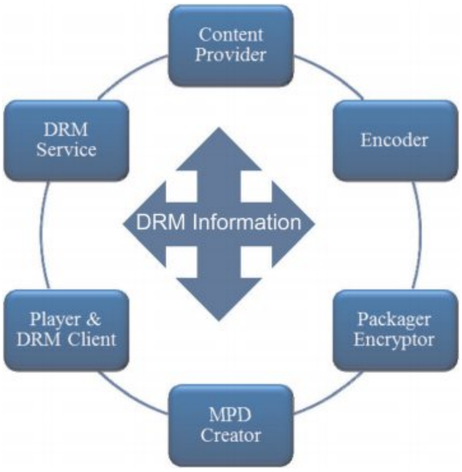


Figure 1 *Logical roles that exchange DRM information and media.*

Packager/Encryptor – A service provider who encrypts and packages media files, inserting default_KID in the file header 'tenc' box, initialization vectors and subsample byte ranges in track fragments indexed by 'saio' and 'saiz' boxes, and possibly packages 'pssh' boxes containing license acquisition information in the file header. Tracks that are partially encrypted or encrypted with multiple keys require sample to group boxes and sample group description boxes in each track fragment to associate different KIDs to groups of samples. The Packager could originate values for KIDs, content keys, encryption layout, etc., and then send that information to other entities that need it, including the DRM Provider and Streamer, and probably the Content Provider. Alternatively, the Packager could receive that information from another entity, such as the Content Provider or DRM Provider.

MPD Creator – The MPD Creator is assumed to create one or more types of DASH MPD. The MPD must include descriptors for Common Encryption and DRM key management systems, and should include identification of the default_KID for each AdaptationSet element, and sufficient information in UUID ContentProtection Descriptor elements to acquire a DRM license. The default_KID is available from the Packager and any other role that created it, and the DRM specific information is available from the DRM Provider.

DRM Client – A player typically relies on a native DRM client installed on a device that must receive information from different sources: MPD, Media files and DRM licenses.

DRM Service – The DRM Provider creates licenses containing a protected content key and playback rules that can only be decrypted by a trusted client.

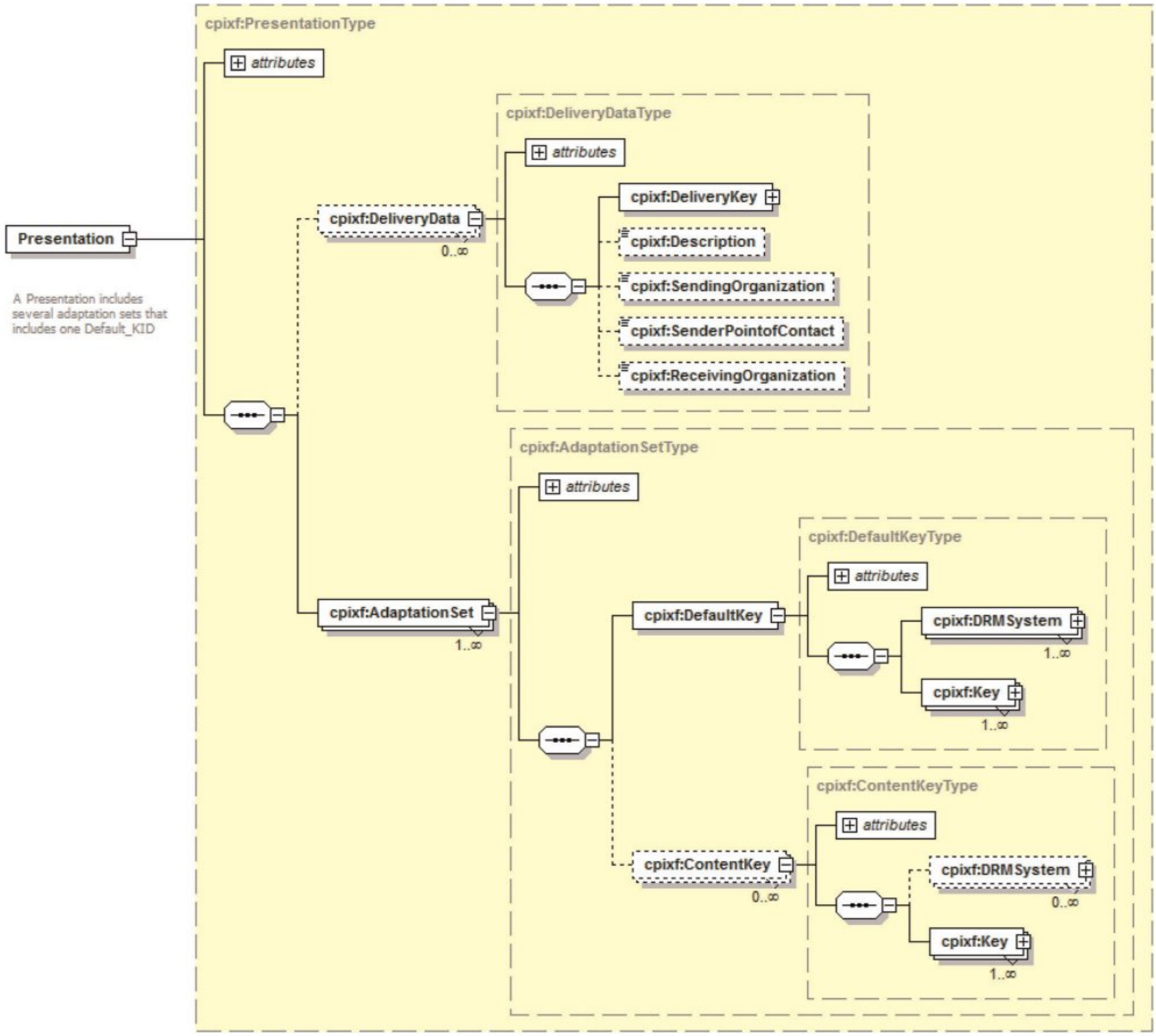


Figure 2 *The CPIXF high level view.*

The DRM Service needs to know the default_KID and DRM SystemID and possibly other information like asset ID and player domain ID in order to create and download one or more licenses required for a DASH presentation on a particular device. Each DRM system has different license acquisition information, a slightly different license acquisition protocol, and a different license format with different playback rules, output rules, revocation and renewal system, etc. The DRM Service typically must supply the Streamer and the Packager license acquisition information for each UUID ContentProtection Descriptor element or 'pssh' box, respectively.

The DRM Service may also provide logic to manage key rotation, DRM domain management, revocation and renewal and other content protection related features.

In such ecosystem, there can be different content preparation and information workflows, therefore CPIXF uses a container that is similar in structure to an MPD to allow secure exchange of all DRM information between any entities in any workflow.

DASH-IF has defined a container called the Content Protection Exchange Format (CPIXF) which has the following main properties. This is an XML file that is fully described in [DASH-CPIXF].

The Content Protection Exchange Format

The structure is similar to the MPD structure defined in [DASH]. A Presentation is the root element of this schema and contains all information required for getting the common encryption keys which is used to encrypt all representations within all adaptation sets. It follows these principles:

- Following the constraints defined by [DASH-IOP], it is assumed that the same key is used for encrypting all Representations of a given Adaptation Set. For supporting key rotation, several Content Keys can be used for encrypting all Representations, each key with a validity period.

- The same XML file can be shared between several receiving entities; hence, each one must be able to decrypt the encrypted Common Encryption keys contained in the document by using public and private keys shared with the sender. The sharing of public key pairs is usually part of a contractual relationship between entities authorizing access to the content and keys, and is outside the scope of CPIXF.

- Taking this into account, the Presentation contains:

  - DeliveryData: Each instance of the DeliveryData describes an entity that is permitted to decrypt common content key contained in the XML document. There is textual description and associated certificates for example.

  - AdaptationSet: Each AdaptationSet contains the DefaultKey information (the common content key itself and all associated DRM Signalizations which is protection system specific information for every DRM.). Optionally, it can also contain ContentKey instances used when key rotation is enabled on this Adaptation Set.

The keys inside the DefaultKey and ContentKey entities can be encrypted inside the XML file using information provided in the DeliveryData element. The XML file also allows storing the content keys in the clear and then the protection of the delivery mechanism, such as IPSEC or TLS, is used for securely deliver the file.

The proposed schema relies on the Portable Symmetric Key Container (PSKC) defined by IETF [RFC-6030] for describing keys and the associated encryption. Necessary extensions are added for supporting DRM information and association of information with content.

Use cases

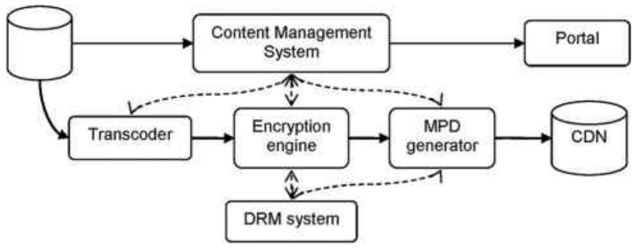
The following describes two classical cases where the CPIXF helps the overall workflow implementation. For on-demand content and live content,

Content on demand

The flow for preparing Content on Demand requires that a file is available non-encrypted, ideally in the maximum resolution so that DASH content can be prepared.

A Content Management System (CMS) masters the creation flow. The CMS makes the file available to a transcoder. The transcoder outputs the segmented files that can be encrypted. The encryption engine either generates the content key(s) or requests them from a DRM system. The DRM system also provides any information to be added in the PSSH boxes. When the encrypted DASH content is ready, the MPD is generated by a MPD generator. It asks the DRM system the required DRM signalization to be added in the MPD. DASH content is then uploaded by the CMS on a CDN making it available to users. In parallel, editorial metadata is exported to the Portal, enabling access to users. DRM systems receive relevant metadata information that needs to be included in the license (output controls) when creating a license.

This flow is summarized in the following figure where arrows show the flow of information.



In this flow, the CPIXF finds its natural place in all exchanges between the DRM system and Encryption engine and between the DRM system and MPD generator.

Electronic sell-through

In order to make available its content in a defined and controlled quality, a Content Provider is preparing it. Preparation includes transcoding to the desired format and encryption of the resulting segments. The content owner is generating also the content key(s). At the end of the process, DASH content is ready and stored along with the content key(s).

Later the content owner distributes the prepared content to multiple locations with the addition of metadata describing it. Content becomes then saleable on multiples Portals. In parallel, the Content Provider distributes the content key(s) to any authorized DRM system. A DRM system is authorized if it is one used by one of the Portal that has this content for sale.

In this flow, the CPIXF finds its natural place in all export of content key(s) from the Content Provider to the DRM systems. The Content Provider could also use the CPIXF for securely store content keys along with content.

SD, HD, and UHD content

A Content Provider typically controls keys and license policy for a Presentation, and requires some type of client authentication (is playback requested by a subscriber?) and content authorization (does that subscriber have rights to the requested content and license?). For instance, a purchase or subscription might only include access to high definition content (HD), because ultra-high definition (UHD) content is sold at a higher price. Rights may also be limited by rental period, by location, by device, etc.

SD, HD, and UHD Adaptation Sets may require different keys and licenses because each requires a different security level and output controls. For instance, it is typical to allow SD content to be output over analog interfaces, but restrict HD content to protected digital outputs such as HDMI with HDCP. HD content may also require separate keys for audio and video because audio keys are typically less protected in devices. UHD content may require a hardware protected video path and HDCP 2.2 output protection.

A Content Provider can enable these scenarios by determining the mapping of KIDs to Adaptation Sets in CPIXF, then specifying the license policy required for each KID with DRM license Providers. Each player should determine in advance what content it is entitled to, and what level of content protection it supports; and then request the necessary licenses before attempting playback. A request will usually be authenticated and authorized by the service provider according to their business rules before providing the client an access token it can use to request the type of license authorized. The DRM license provides cryptographic enforcement of the entitlement authorized.

Separate from CPIXF, there is a contractual relationship of rights and responsibilities flowing from the copyright holder to each entity that handles the content and keys. Entities that handle unencrypted content or keys are contractually required to protect them, and are typically given cryptographic certificates containing the keys necessary to send and receive media keys protected by CPIXF key encryption.

Conclusion

This paper presented a secure mechanism for exchanging sensitive multimedia information when preparing value-added content. It makes no assumption on the overall trust framework and supports several use cases, from simple ones to more complex.

It provides an additional step in interoperability when enabling protected content with DRMs.

In next steps, the proponents will propose extensions allowing to fully supporting common use cases, such as on-demand content and live content. The first one is a quite straight forward use of the CPIXF while the latter requires the management of key rotation.

Acknowledgements

The authors would like to thanks the DASH-IF forum for its support in conducting this work and more specifically, all attendees to the DRM group who actively participated in the specification of this implementation guideline.

References

[1] [DASH-CPIXF] DASH-IF, 2015. Implementation Guidelines: Content Protection Information Exchange Format, March

[2] [DASH] ISO/IEC 23009-2:2014. Information Technology – Dynamic Adaptive Streaming over HTTP (DASH) – Part 1: Media Presentation Description and Segment Formats

[3] [DASH-IOP], 2014. DASH-IF Guidelines for Implementation: DASH264/AVC Interoperability Points, August

[4] [RFC6030] IETF RFC 6030, 2010. Portable Symmetric Key Container (PSKC), October

Selected content from IBC2015

4G Broadcast: can LTE eMBMS help address the demand for mobile video?

A.J. Murphy   C.R. Nokes

BBC Research & Development, United Kingdom

Abstract: At a time of increasing demand for video on mobile devices, it is vital to use the resources on existing mobile networks as efficiently as possible.

4G Broadcast (LTE eMBMS) offers the possibility of addressing issues of congestion and peak demand for popular content by sending a single stream once for reception by multiple users within a cell. This capability could be enabled in localised areas of peak demand, within existing 4G networks, and switched on as required, to allow the network to be dynamically optimised for the current traffic conditions.

BBC Research & Development has been investigating how 4G Broadcast technology might be used to improve the delivery of streamed content to mobile devices. We have demonstrated two example use cases – firstly as part of an app tailored to a specific event (for example at a sports venue), and secondly by connecting the technology seamlessly to BBC iPlayer (the BBC's Internet streaming service that offers both live and catch-up content) to allow viewers to continue watching popular content in congested areas, without the experience being spoilt by buffering.

This paper will explain the work BBC Research & Development has been carrying out on 4G Broadcast and present the results of recent trial work.

Introduction

In recent years, the proliferation of highly capable smartphones means many more people are now interested in watching video on their mobile phones, and there is therefore a corresponding increase in the amount of content available for such devices, both as short-form clips as well as long-form programmes and live streams. Although much of this viewing is done whilst the device is connected to a WiFi network (either at home or in the office etc.), there is also demand for this content directly over the mobile broadband networks – and the user is probably agnostic to the nature of the connection, being more concerned about both reliability and cost (or limits from data allowances).

Most industry analysts expect that the demand for mobile data will continue to increase significantly, and that a large portion of this will be driven by demand for video. The concept of the "busy hour" is already well known within the industry when mobile networks are at their most congested at places of peak demand (e.g. busy railway stations during commuting hours), and increasing numbers of people requesting popular video content can only exacerbate this situation. There is clear evidence that, for example during key sporting events such as the Wimbledon tennis championship, demand for live video content to mobile devices has increased dramatically; on 7th July 2013, when Andy Murray became the first British man to win the Wimbledon title since 1936, 64% of total requests to the BBC Sport site were from handheld devices (1). 4G Broadcast also has the potential to address spikes in demand over the mobile network when a new show or film is first released.

In all these cases, where large numbers of people are trying to watch the same content, at the same time, within the same geographic area, using a mobile broadband network, currently the network will attempt to stream numerous identical copies of the content, using a unicast mechanism. 4G Broadcast offers the opportunity to replace this with just one single stream for all users in one area, which not only has the potential to be significantly more efficient, but should also allow all users to be assured a consistent good quality experience, without the risk of buffering due to network congestion (in areas of good coverage).

This paper concentrates on the technical aspects of 4G Broadcast and does not address wider business or cost issues (for example the impact of 4G Broadcast on data allowances) or rights considerations.

What is 4G Broadcast?

The term '4G Broadcast' is used within this paper to refer to Long Term Evolution (LTE) enhanced Multimedia Broadcast/Multicast Service (eMBMS), the broadcast mode defined by the 3rd Generation Partnership Project

(3GPP) in their 4G standards. Use of the term means eMBMS, as currently specified, namely within 3GPP Releases 9, 10 and 11.

Choosing this term was done for two reasons; firstly the aim was to use a name that would be accessible to a wider, non-technical audience; secondly other commonly used terms within industry such as ‘LTE Broadcast’ are not always used consistently and are sometimes used to refer to potential future developments of the specification.

There is typically benefit to be had from the broadcast mode when 2-3 users want to watch the same content concurrently, although the exact point at which it is more efficient to switch from unicast to broadcast will be dependent on the precise propagation conditions and how far the users are from a given cell tower.

Technology

eMBMS (2) shares the physical layer OFDM modulation scheme with LTE, allowing a certain proportion (up to 60%) of the LTE sub-frames to be assigned to the broadcast mode. These sub-frames are designed such that they do not adversely impact handsets that do not support them.

The symbols of these frames are defined to use a longer Guard Interval than conventional LTE, allowing a path difference of up to around 5 km within a synchronised Single Frequency Network (SFN) across multiple cells – another potential source of efficiency gain. As such, eMBMS is primarily targeting delivery to mobile devices and is therefore not suitable as a replacement for systems such as DVB-T (3) or DVB-T2 (4).

Whereas with conventional unicast LTE, a handset has the option to request retransmission of data it is unable to decode, the broadcast mode is a one-way transmission. As a result, and to add time diversity, an additional Application Layer Forward Error Correction (AL-FEC) scheme is added in the form of Raptor codes.

MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH) is the dominant delivery format over eMBMS. It is a segmented video format that splits the video/audio stream into a number of chunks which are delivered as individual files. These files have unique filenames which are reassembled at the receiver into a continuous stream. A manifest file is used to signal to the receiver the video and audio format used and the naming convention for the segment filenames. MPEG-DASH is particularly suited to Internet delivery where it can traverse firewalls in the same way as normal web traffic and lends itself to caching and efficient delivery over Content Delivery Networks (CDNs). It is also adaptive with the client able to choose from a number of different representations at different bit-rates as the network connection dictates.

It should be noted that for use by eMBMS, DASH is not in fact used in a dynamic or adaptive fashion, nor is it delivered over HTTP. Instead, a single representation is delivered over a fixed bitrate broadcast bearer using a data carousel.

For delivery over eMBMS, the individual video and audio segments are packaged into a data carousel using the FLUTE protocol. This packaging, and the application of the AL-FEC, is the task of a new entity within the mobile core network called the Broadcast Multicast Service Centre (BM-SC)

In terms of media delivery, both streaming and file transfer are defined within eMBMS. However, the use of MPEG-DASH over the file transfer mechanism appears to dominate and has the advantage of integrating well with unicast delivery.

How could 4G Broadcast be useful?

Figure 1 shows situations where eMBMS might be useful. The vertical axis shows the two main categories of consumption; linear/live programmes vs. on-demand. The horizontal access indicates location. If a user is at home and wishes to consume either form of content it is likely that they will be able to stream over a local WiFi connection (on the assumption that a sufficiently good fixed broadband connection is available).

Away from home and with on-demand content, a user has the option of downloading in advance before leaving home (side-loading) or streaming over conventional 3G/4G networks where data allowances and capacity allow.

Linear or live content, by definition needs to be consumed now, whether it be a live sports event or on-demand content that has a linear-type consumption pattern such as demand immediately after a popular new show is released on-line. It is in this area where 4G Broadcast could best play a role in delivering this content most efficiently and with the best quality to mobile devices.

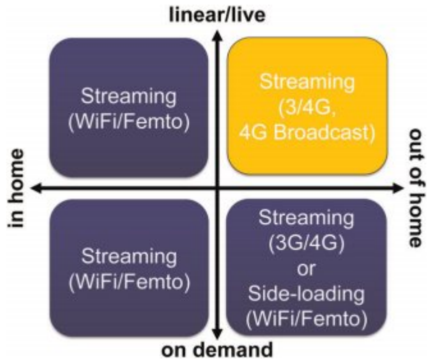


Figure 1 *Application Areas for Delivery to Mobile Devices*

The Commonwealth Games trial

In the summer of 2014, a trial of 4G Broadcast was carried out as part of BBC R&D's wider public showcase around the Commonwealth Games at the Glasgow Science Centre.

The trial was a collaboration, with BBC R&D providing content and an application, EE providing a network and dedicated spectrum, Huawei supplying equipment and Qualcomm providing software and middleware. The handsets were Galaxy S5s supplied by Samsung. These were off-the-shelf handsets (since they already support eMBMS within the hardware) but with dedicated firmware to enable reception of the broadcast streams.

A 2.6 GHz frequency allocation with a 15 MHz bandwidth was used to provide a private LTE network from a dedicated eNodeB (base station) transmitting the broadcast signals within the confines of the exhibition hall.

Three streams were made available of the BBC's TV channels carrying live action from The Games. The target screen size was 5.1" which meant that standard definition resolution was sufficient and an average video bitrate of 1.3 Mbit/s was used with an MPEG-DASH segment length of 1s. The use of short segments reduces the impact of error extension and ensures that, in the event a segment cannot be recovered at the receiver, the impact to the viewer is minimised.

Trial architecture

Figure 2 shows the architecture put in place for the trial.

The flow of the audio/video content through the system begins at BBC Centre House in West London, where direct feeds from the BBC's playout centre were encoded and formatted as MPEG-DASH streams. These were then transported to EE's test lab, via Telehouse in London over a dedicated link put in place for the trial. Here the BM-SC carried out the eMBMS encapsulation as well as applying the AL-FEC. A Packet Data Network Gateway (P-GW) was also present to allow conventional unicast Internet access on the trial LTE network.

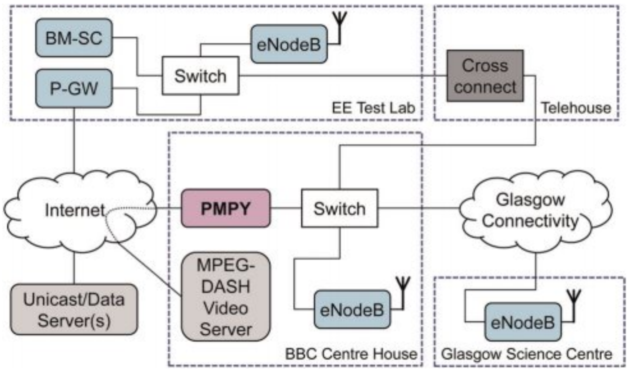


Figure 2 *The Architecture (simplified) for the 4G Broadcast Trial*

From the EE Test Lab, the encapsulated content then returned to BBC Centre House, before being sent up to the Glasgow Science Centre via the dedicated connectivity put into place for the showcase for transmission within the exhibition hall.

As well as an eNodeB within the Glasgow Science Centre, there were also eNodeBs present within the EE Test Lab and at BBC Centre House to enable detailed testing to be carried out.

One particular node of note in the diagram is the 'PMPY' or Push Me Pull You. The task of this was to act as an interface between a standard MPEG-DASH server and the BM-SC. MPEG-DASH streaming is normally driven by the client, which, based on the current time of day, requests or *pulls* a segment with a particular index sitting on the HTTP server. However, BM-SC implementations typically expect segments to be *pushed* to them as soon as they are available. The task of the PMPY was therefore to act as a conventional MPEG-DASH client to *pull* segments from the server and *push* them to the BM-SC. In practice this consisted of a computer running the Linux operating system and some software developed specifically for the trial. The PMPY also served as a useful monitoring and logging point for the duration of trial.

The user experience

Two potential ways of utilising 4G Broadcast were presented to the public. The first was a dedicated events-based application specific to the Commonwealth Games as shown in Figure 3 below.

This allowed the user to navigate around the different sports venues using a map with pins appearing green when a live stream was available from that location. Clicking on the venue brought up information about the event currently in progress and allowed the user to see the live feed. In this application of the technology, users would be aware that in some sense the 4G Broadcast enabled a special service for an event.

It was necessary to bring a number of data feeds together to drive the application as Figure 4 shows. Data concerning the venue names and locations, the sports events taking place at each one, the TV schedule and the mapping of TV channels onto those being broadcast over eMBMS were all combined into a single file using software written in Ruby and stored on a web server for the application to interrogate.

The second approach was to show eMBMS as an underlying technology that could enhance reception of mobile video on an existing service. To demonstrate this, the 4G Broadcast streams were connected to the BBC

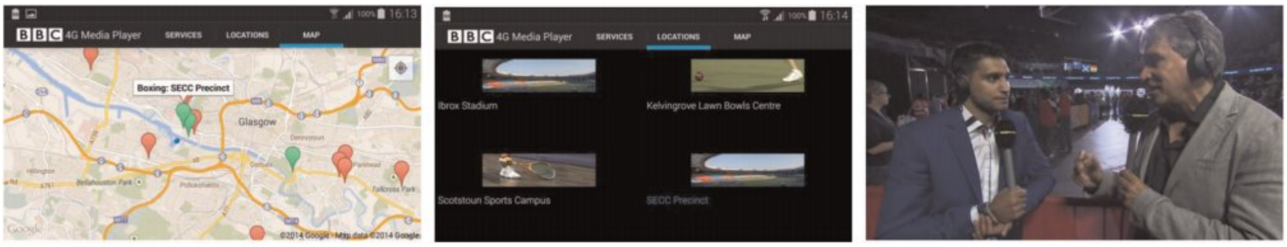


Figure 3 *The 4G Broadcast Application*

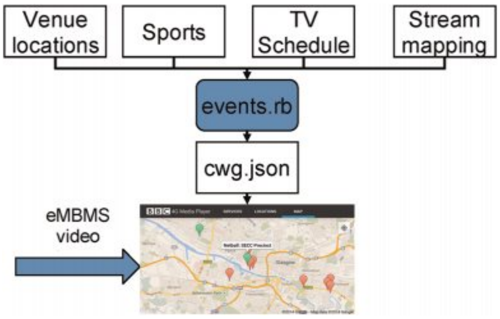


Figure 4 *Data Sources for the Application*

iPlayer, the BBC's Internet streaming service that offers both live and catch-up content. This is shown in Figure 5. In this way, the user was taken seamlessly to the 4G Broadcast stream when it was available while still having catch-up content available over unicast streaming.

Lessons learnt

Public reaction to the demonstration was very positive with many praising the quality and consistency of the 4G Broadcast pictures. There was also a good awareness of issues such as data caps and congestion that can effect unicast streaming on current mobile networks in certain situations and therefore the potential advantages of 4G Broadcast.

Early on in the project, we realised the need for the PMPY entity to act as the interface between the mobile network and BBC R&D's standard MPEG-DASH servers. Unlike the vast majority of interfaces within 3GPP, this interconnection is not currently specified and, although it seems that most manufacturers use similar protocols, this is something that could perhaps benefit from standardisation within 3GPP in order to simplify integration with different operators and vendors.

Another potential issue with the current eMBMS standard is the lack of support for statistical multiplexing. On conventional broadcast services, the BBC shares bitrate across multiple services. This means that if there is content on one channel that is simpler to encode, some instantaneous capacity can be released for another channel to use in the event that it is showing something that is more challenging. This approach results in an improved overall picture quality.

Figure 6 above is indicative of the sort of variation in bit-rate seen from one configuration of the MPEG-DASH encoder used for the trial over the course of around 15 minutes. While it would be possible to constrain this further, it would result in a reduction in picture quality. Since the broadcast bearer is a fixed bitrate pipe specified per service, there are two main options. The first is to over-specify the bit-rate to cope with peaks in capacity; the second is to accept increased delay through the system. Neither of these is optimal and support for statistical multiplexing across services is something that could be very beneficial.

Finally, the end-to-end delay through the system is something that needs further characterisation. Since MPEG-DASH relies on a sequence of files, there is the risk that the presence of buffering at multiple handover points through the chain serves to introduce significant delay. This is not so much of an issue away from a sporting event but could impact the use of the technology to show live video of an event inside the stadium where the images on screen could lag behind the live action at the event.

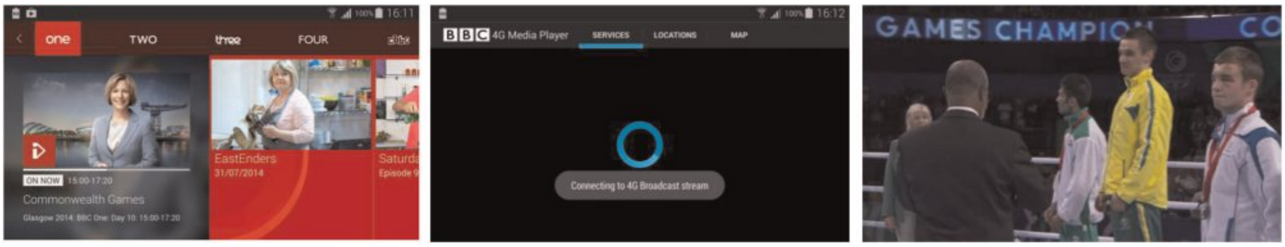


Figure 5 *4G Broadcast and the BBC iPlayer*

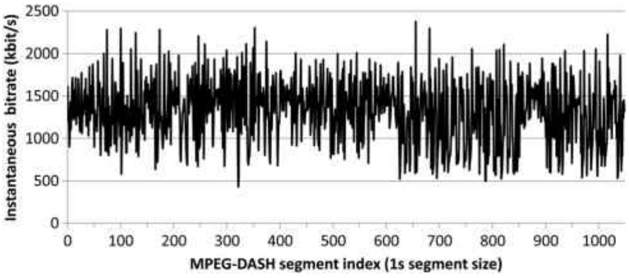


Figure 6 *An Example of the Variation in MPEG-DASH Segment Sizes*

Seamless switching and hybrid networks

The ability to switch seamlessly between broadcast and unicast streams would mean that a user could benefit from 4G Broadcast automatically when moving into areas where it is available. BBC R&D has investigated whether this switching could be carried out in the application, without the need for standardisation.

The use of MPEG-DASH simplifies the implementation of this switching since each segment is uniquely identified by its number. At any given time, the client can calculate the expected segment number based on the difference between the current time and the known start time of the stream divided by the segment duration. Introducing some buffering allows the application to determine if a given segment is received correctly over the broadcast mechanism and, if not, it can attempt retrieval over a unicast connection before it is required to be played out. This is illustrated in Figure 7 below. Inevitably this type of approach introduces some delay but was demonstrated to be work successfully.

The 3GPP Release 12 specification is to include a feature known as MBMS Operation On Demand (MooD). As well as offering a standardised means of implementing this switching functionality, it will allow the network to determine the number of users receiving a particular stream over unicast. Once this reaches a given threshold, the users can automatically be moved over to a broadcast version of the same content, effectively offloading the multiple requests onto a single multicast stream. This would ensure that the resources of the mobile network are always used in the most efficient way possible.

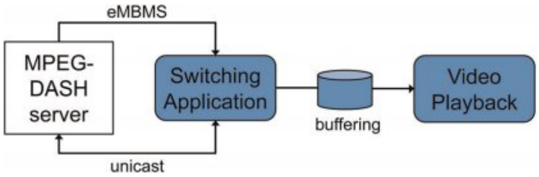


Figure 7 *Application-based Seamless Switching*

The availability of these types of features, enables a hybrid, dynamic network to be built where eMBMS is deployed in the operator's busiest areas and enabled for the most appropriate content at the most appropriate time.

It is this flexibility that has the potential to differentiate 4G Broadcast from previous mobile TV standards such as DVB-H, DMB or MediaFlo that required a new network to be built everywhere that coverage was required.

4G broadcast at the FA Cup Final

Working with the partners from the Commonwealth Games, with the addition of EVS and Intellicore, we participated in a trial at the FA Cup Final at Wembley Stadium in London in May 2015.

As well as the live BBC One feed of the Cup Final, we used BBC R&D's Stagebox technology to deliver an additional two live camera feeds from the outside broadcast area via an IP link. These were then broadcast along with highlights packages. In addition, EVS supplied multi-angle replays to a 'replay zone', allowing users to interactively select the angle of view for incidents of interest during the game.

All of this content, with the addition of real-time statistics, was brought together in a dedicated application written by Intellicore. This ran on a number of specially enabled tablets given out to invited guests.

BBC R&D also worked with BBC Sport to seamlessly connect the broadcast streams into the live coverage area within a modified version of the existing BBC Sport application, shown in Figure 8.

This proved the ability of 4G Broadcast to deliver high-quality content in situations where congestion might not typically allow it and showed the benefits for the user in bringing unicast and broadcast content together in a seamless fashion to give the best possible experience.

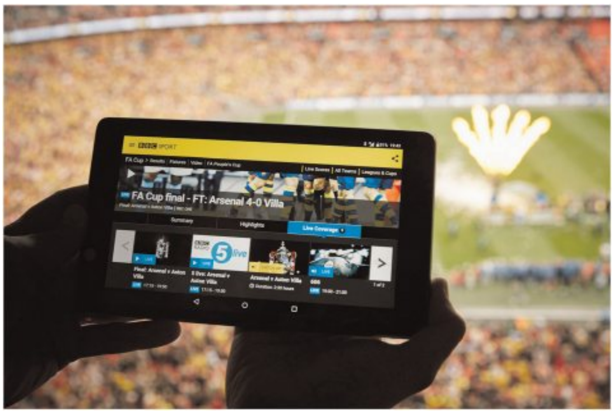


Figure 8 *The BBC Sport App at the Cup Final*

What next?

This paper has specifically concentrated on today's 4G Broadcast, and we would expect to continue to work with the mobile operators and others in the industry in the future to explore further how best to make use of this technology, and whether to support more frequent use at appropriate events.

However, it is also worth considering how the technology itself could be enhanced. In addition to some further developments already being planned for current releases of 3GPP standards (e.g. MooD), discussions have started within 3GPP about the possibility of developing eMBMS technology further. Improvements being considered include allowing sharing of a single broadcast across multiple mobile operators, perhaps making full use of a standalone block of spectrum, as well as introducing longer Guard Intervals to improve coverage across larger areas. It is not yet clear whether some of these features will be fully incorporated within further releases of 4G specifications, or whether some of them will need to wait for the next generation mobile broadband specification (5G), for which the requirements are now beginning to be captured.

One requirement that could be useful, and significantly help both improve the overall efficiency of use of the network, as well as providing for a better experience for viewers, would be the ability of a service operator to use the broadcast mechanism to push popular content to be transparently cached on devices, where this is enabled by the user. This could be done, for example, at times of the day when the network is lightly loaded.

As well as improving the physical layer capabilities, a key requirement from a broadcaster's perspective is that 5G technology should allow for free-to-air delivery (for example of a public service broadcaster's content), without the viewer incurring additional costs.

BBC [R&D] are members of the 5G Innovation Centre at the University of Surrey, and will be contributing to discussions about requirements for 5G technology, to ensure that broadcasters' needs in this area are represented.

Conclusions

Our trial work indicates that 4G Broadcast is capable of delivering a high standard of video and audio to mobile devices with a defined quality of experience, even in crowded environments. Bringing broadcast and unicast together in a single application enables the user to benefit from the strengths of both distribution techniques, while the flexibility to switch seamlessly between broadcast and unicast could allow 4G Broadcast to form part of a hybrid network with it enabled at the busiest locations at the busiest times.

Looking to the future and towards 5G, we have learnt that a number of features could be added or enhanced, such as the introduction of support for free-to-air and stand-alone transmissions, statistical multiplexing between services and more flexible scheduling of broadcasts combined with the ability to seamlessly switch between broadcast and unicast.

Despite the potential of perceived 'infinite' capacity from 5G standards, the realities of a practical rollout are likely to mean that such capability will not be available everywhere and a properly integrated, flexible broadcast mode is therefore likely to be an important factor in delivering the high quality content that users demand wherever they happen to be.

References

[1] http://www.bbc.co.uk/mediacentre/latestnews/2013/bbc-wimbledon-stats

[2] ETSI TS 136 211. LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation

[3] ETSI EN 300 744. Framing structure, channel coding and modulation for digital terrestrial television

[4] ETSI EN 302 755. Frame structure channel coding and modulation for a second generation digital terrestrial television broadcasting system (DVB-T2)

Selected content from IBC2015

A framework for a context-based hybrid content radio

Paolo Casagranda[1] *Alexander Erk*[2] *Sean O'Halpin*[3] *Dominik Born*[4] *Walter Huijten*[5]

[1]*Rai – Radiotelevisione Italiana, Centre for Research and Technology Innovation, Italy*
[2]*IRT – Institut für Rundfunktechnik, Germany*
[3]*BBC – British Broadcasting Corporation, United Kingdom*
[4]*TPC AG, Switzerland*
[5]*NPO - Nederlandse Publieke Omroep, Netherlands*

Abstract: The aim of this paper is to propose hybrid content radio, a new framework for radio content, enhancing the traditional broadcast radio experience and augmenting it with context related audio content. Differently from most of the commercial recommendation-based internet streaming services (Spotify, Pandora), here we consider systematically adding audio content to an existing, linear audio structure. The purpose of the hybrid content radio framework is to enhance the broadcaster's programme schedule with context-aware and personalised audio content from the internet. The context can be the listener's profile, emotional state and activity, their geographical position, the weather, and all factors contributing to characterize the state of the listener. The final purpose of the enhancement is to improve the service user's' listening experience, decreasing their propensity to channel-surf and giving them more targeted content, such as local news, entertainment, music and also relevant advertisements. In this way, the hybrid content radio approach enables both a functional enhancement to radio and network resource optimization, allowing the use of both the broadcast channel and the internet.

Introduction

This paper gives an overview of the recent experimental services proposed by a group of European Broadcasters exploring the potentialities of a hybrid approach for audio in radio. *In hybrid content radio (HCR), traditional linear broadcast radio is the foundation upon which a new, enriched service is built, using enriching audio content from the broadcaster's archives or from trusted third party providers*. The paper presents experimental services and outlines key technical requirements for the creation of an HCR radio framework. Different from most of the internet streaming services, here we consider adding audio content to an existing, linear audio structure: the broadcaster's programme schedule. Specifically, HCR allows enhancement of the broadcaster's linear schedule with context-aware and personalised audio content. The context can be the listener's emotional state, her geographical position, her group, the weather and all factors contributing to it [1]. The final purpose of the enhancement is to improve the service user's listening experience, giving her more targeted contents, such as news, entertainment, music and also relevant advertisements. The proposed technique achieves content personalisation at a minimal bandwidth cost, as the broadcast channel is used if possible, differently from existing internet music playlists. In this way, HCR allows an optimized bandwidth usage. Figure 1 illustrates the concept: broadcast linear audio content is enriched by personalised content from the internet.

The proposed framework can be applied to both audio and video content. However, audio is well suited as a background medium, and can be enjoyed while people are doing something else. It's common to see people listening to radio while walking, biking or driving or engaged in different activities. In this sense, context has a more complex impact on radio than on television.

Related work

There are already several mobile music streaming services creating highly personalised playlists, exploiting different content recommendation techniques: Pandora, based on content features extracted by experts, the Music Genome Project [2], others mainly based on collaborative filtering or hybrid techniques like Spotify [3]. Music streaming services generally use recommender systems exploiting collaborative filtering, content-based or social-based techniques [4-6] and exclusively use the internet channel to reach listeners with wholly customized playlists. Different from those

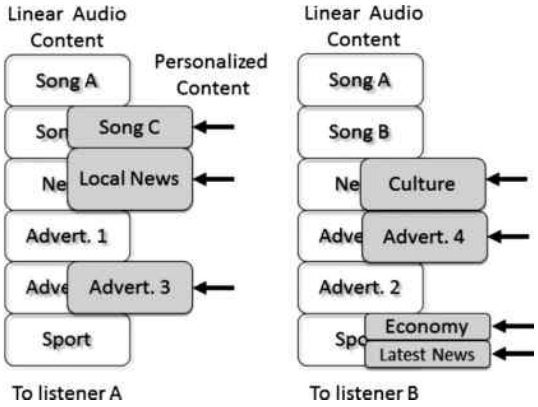


Figure 1 *The hybrid content radio concept: broadcast linear audio enhancement by audio content replacement*

services, *hybrid content radio addresses all the scenarios where the linear audio content is partly and flexibly replaced by personalised audio content, and the broadcaster maintains overall control.* Several IST European Projects have considered enhancing broadcast video content, recombining broadcast objects at the client and downloading them from the broadband network. Specifically, the iMedia and Savant IST Projects [7-9] used targeted, personalised advertisements or recommended content lists for television services; however context-awareness was not considered. In past years, music and video recommenders generally suggested personalised playlists and dynamic content only focusing on user or item-similarity, the customer profile and navigation history. In contrast, recent research and commercial services have started addressing contextual information, leveraging the context such as the user's position, mood or activity [10]. Most recently, a number of studies analysed the proposition of context-aware audio content. Contents and services related to the listener's location are emerging, like the Foxtrot prototype [11], a mobile location-based and crowd-sourced audio application playing an automatically generated playlist from geo-tagged music. Other examples are MIT BusBuzz [12], creating a social music experience while on bus, and the MIT Loco-radio Project [13], a mobile, augmented-reality system creating an audio landscape while on the move. HCR maintains the broadcast schedule as the framework to build a new service upon, also considering it as part of the context, so that the enhancing audio content has to be included in accordance with the schedule. This is a novel element compared with previous studies. HCR has its roots in the work carried out by the RadioDNS project. RadioDNS has addressed hybrid radio, focusing more on the enrichment of audio content with images, text and metadata, and also allowing to link external content, see the ETSI technical standards [14].

Proof of concept

The following sections will give an overview of the experimental HCR services developed by partners Rai, IRT, TPC and NPO.

GeoRadio: geo-referenced hybrid content radio

Rai has developed an HCR proof-of-concept based on contextual information, leveraging geo-referenced audio, called "GeoRadio". The listener's context is the primary source of information to create a more personalised radio experience. In particular, the user's position and destination play key roles in making better content recommendations (see Figure 2). Position information has already been used by other audio streaming systems, specially audio-based city guides, geographically and POI (Point of Interest)-based music, ambient sounds, audio information and user-generated content [11]. The GeoRadio prototype is the first demonstrator developed by Rai to show the potential advantages of adopting a more flexible framework for radio. This service uses the listener's position and destination to select and recommend part of the audio content. A simple scenario will give a quick overview of how the prototype works. The listener, while driving in Turin, is listening to the radio with her smartphone, and the app is tuned on Radio2 channel. When the national news program is about to finish, the recommender proposes the regional news (and the app connects to the internet to get the audio clip from the Rai servers). Then, when approaching the old Filadelfia stadium, the radio proposes to her a related historic audio document. During the advertisements break, the app proposes a targeted advertisement: "Amici Miei" restaurant, in the neighborhood - easy to reach at lunchtime, as the recommender keeps track of her habits during the day. In this scenario, the traditional linear radio channel from Rai was enhanced with on-demand content: local news, geo-referenced audio clips and location-based advertisements. In the prototype, listeners connect to the radio station using a radio app. When the app starts, traditional linear radio content is delivered to the app using a broadcast protocol. While delivering linear content, the app should use broadcast delivery channels like DAB+ or FM Radio with devices with built-in radio tuners and a standard radio API. Internet streaming is used as a backup

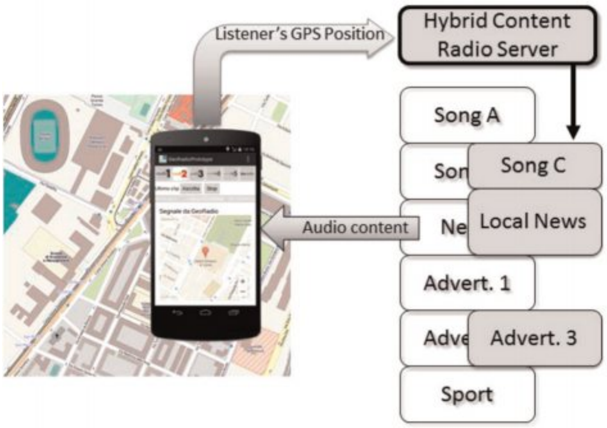


Figure 2 *Audio content recommendations depending on position and itinerary*

solution when broadcast radio is not present. The enrichment process relies on a recommender system suggesting additional, non-linear audio content to be used to replace part of the linear content. At the end, a synchronisation and adaptation component decides the instant in which the additional content is to be played on the device. In this way, the broadcast audio content schedule can be modified in a personalised way for each listener. The possibility of detecting and addressing groups of listeners (e.g. at home, in a car, in a gym) has been analysed in another work [20].

Context based audio: the HbbRadio project

IRT has analysed hybrid content radio within the HbbRadio Project. A first area of investigation targeted dynamic user profiles and context recognition. Novel methods have been developed to understand how to enrich the static radio live stream with personalised elements. These personalised elements are content pieces which are selected to fit the personal preferences of the user and additionally fit into the current context the user finds herself in. In order to achieve this, the first goal is to recognise both the personal preferences and the current usage scenario (context). Based on these findings, a recommendation engine can select suitable content elements and embed these elements into the live stream. One important research topic is the definition of the listening context. The definition of a listener's context is given as: "*A recurring period of time, in which the user is either in the same geographical area or in the same personal activity while interacting with the HbbRadio system*". Therefore the HbbRadio system tracks information such as time, location and actual activity of the user. Figure 4 shows the results of an early stage of the listening situation modeling. The coloured segments show recognized listening situations: blue is work at home, orange is the daily break for lunch, green is the sport once a week and white areas are caused by a lack of power by the device. A second area of interest is personalisation and delivery. Figure 3 shows the components for a general broadcast planning and playout system. All these components communicate via an XML-based interface and exchange planning data and content in real-time. Embedded in this infrastructure, HbbRadio developed a HbbRadio backend component which is responsible for the generation of live data such as SlideShow images, DL/DL+ messages and EPG programme data as well as the provision of archive information for the recommendation engine and on-demand playout services. For the live playout services, a REST API was developed which provides functionality for the HbbRadio backend to request data regarding channels, shows and content of the current schedule. It also provides a callback mechanism for the HbbRadio backend, should changes (e.g. the contents of a single show) be made and need to be propagated into the broadcast. The DAB-EPG Server and DAB-SLSServer components are responsible for gathering the data through the REST API, listen for updates through the callback mechanisms and generate the data formats to be played out over the DAB multiplex. Using a REST API the HbbRadio Engine collects metadata of content which is not scheduled for a live playout, but will be available for the recommendation engine to be integrated into a personalised programme guide. The result is that a user who switches on the HbbRadio functionalities receives a personalised EPG, which is based mainly on the planned live schedule of her favourite station but is enriched with on demand content items at suitable locations. Figure 5 shows the user interface either on the stationary DABerry client or on the mobile Android based HbbRadio client. If personalised content has been inserted into the EPG, the user would see a special icon beside the event. Additionally the HbbRadio client provides a "skip" function for the end user. If the listener is not satisfied with the current audio content, she can skip the audio content to an item from

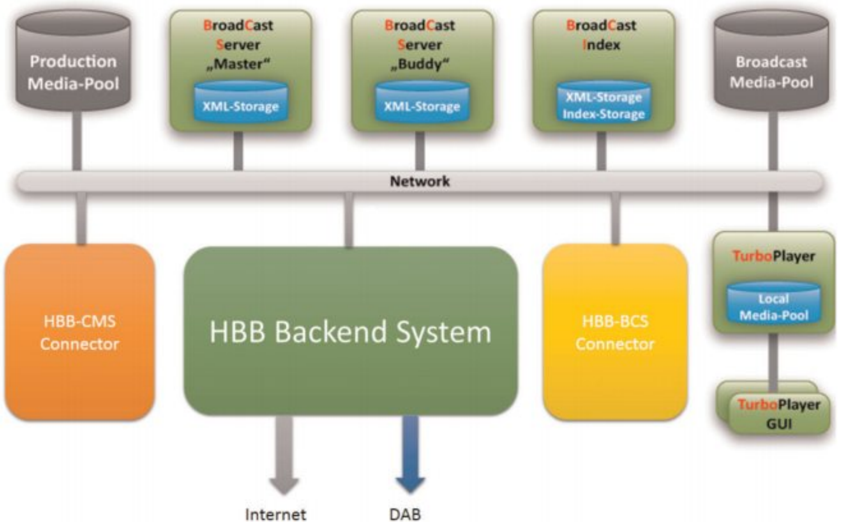


Figure 3 *The Hbb radio architecture scheme*

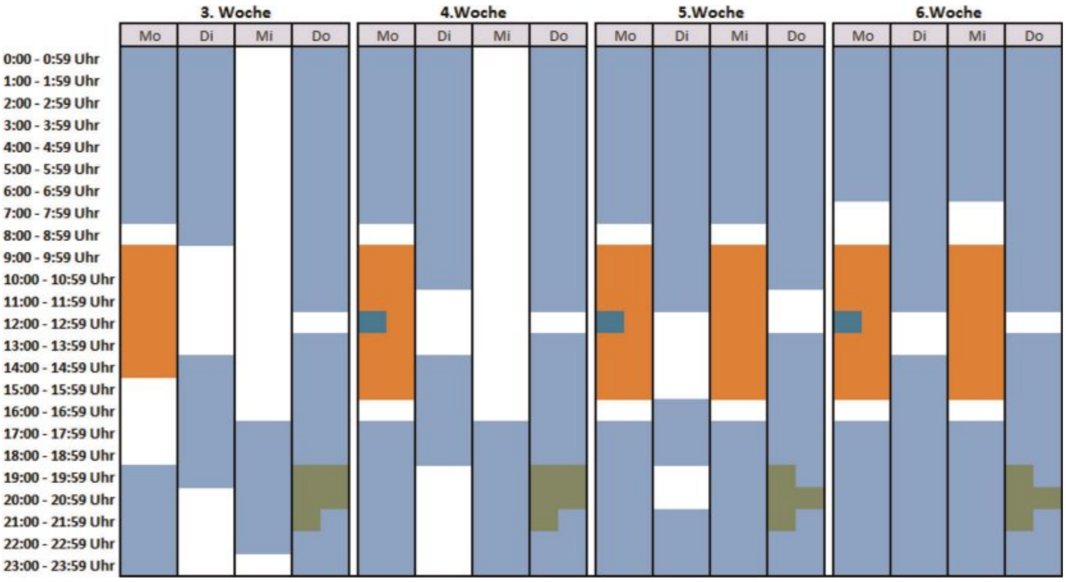


Figure 4 *Example of personalised content replacement during several days*

local storage or alternatively to a proposal from the recommendation engine. The linear programme is recorded into a time-shift buffer and segmented with the ’item running’ and ’item toggle’ signalization in the DAB broadcast stream. These cue points in the time-shift buffer are then re-entry points from on-demand content back into the linear service.

DIY.FM and musicBan: Swiss individual radio

Started as a pilot project during 2012, the diy.fm radio player created by TPC AG combines linear and nonlinear content from the 17 Swiss public broadcaster radio stations. This new radio player allows listeners to create their own personal radio programme by combining linear and non-linear audio content from the Swiss public broadcaster’s (SRG SSR) radio stations with other streams and on-demand services from all over the world. To name one example, one could choose to play the non-stop music programme from "Radio Swiss Pop", but have the player switch automatically to the newsflash of another channel at every full hour. The player also remembers the exact play position, so that a user can change playback devices while listening to on-demand content. When resuming playout on a new device, the on-demand content will continue at the exact position where it was previously stopped. It is now also possible to share the playout position using social networking tools. diy.fm is therefore not only a playout platform, it also provides application programming interface (APIs) for the radios of the Swiss broadcasting corporation. With the diy.fm API, external developers can access the diy.fm content from within their applications or the broadcaster itself can use it for new applications. Diy.fm

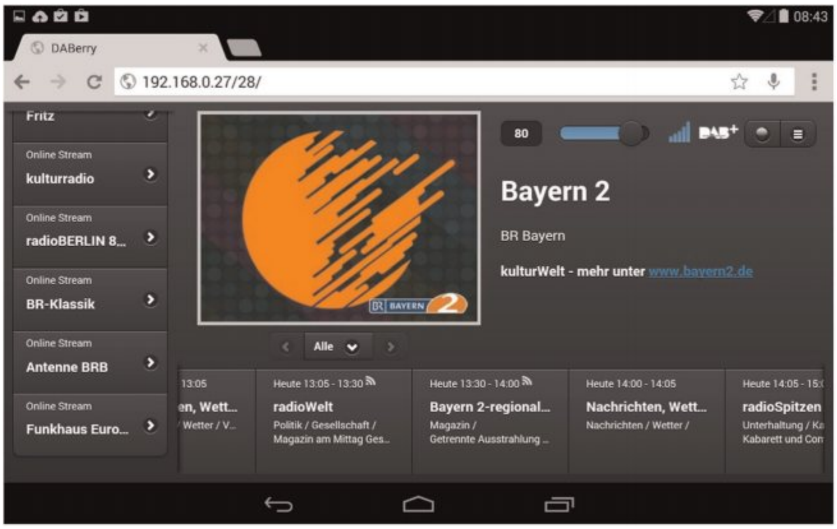


Figure 5 *HbbRadio Interface*

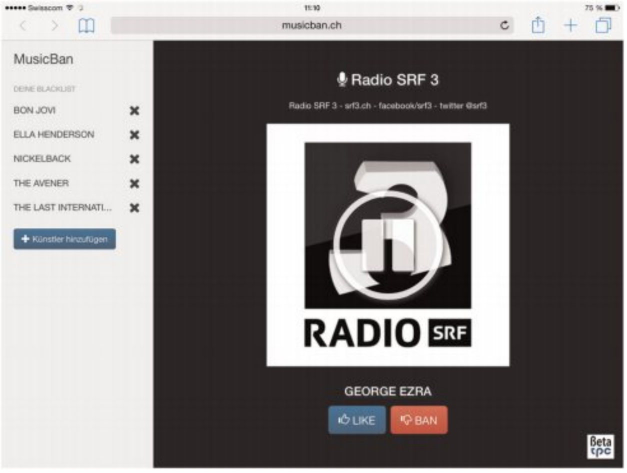


Figure 6 *musicBan web app interface*

allows enriching broadcast content with local music saved on a device. The whole diy.fm API is documented on wiki.diy.fm and is currently being used by internal and external developers to create new applications. The diy.fm APIs consists of a REST API, Java Script Player API and the Push API including News On/OFF, DLS, Traffic announcement On/OFF. The data formats of the APIs are XML, JSON, JSONP. One new prototype based on the diy.fm API is musicBan. MusicBan (Figure 6) is an audio player which combines the diy.fm API and another API from the Swiss music platform mx3.ch. If a user blacklists a song on her player, the song will be replaced by another song with a similar length. The player asks the music platform API for the best song to overlay the blacklisted song. The Dutch broadcaster NPO is working on a similar prototype (code name "Skippy"), proposing another song when the listener indicates she wants to hear an alternative choice from the DJ. This will be implemented in the current 3FM Radio smartphone app but the principles should work for any Radio device, making use of the Cross Platform Authentication (CPA) protocol and the RadioDNS RadioTAG service. The context will initially be limited to the listener's wish to hear an alternative, but other contexts will be considered in subsequent versions.

Guidelines to the creation of a hybrid-content radio service

The previous descriptions of the prototypes demonstrated the common vision for hybrid content radio. The key technical issues towards enabling HCR will be outlined in the following sections, describing a path for further development and research.

Programme metadata and link to enriching content: The first technical requirement for creating an HCR service is to have metadata for the current and future programmes. Metadata must have accurate time and programme details, allowing additional content to be played at the right moment. The recommended specification for metadata is the ETSI Service and Programme Information (SPI) [14]. We propose the following simple mechanism to link the enriching content to the programme schedule. The SPI `<link>` element describes how to link external content: based on that, SPI allows the following syntax:

```
<link    uri="http://broadcaster.example/
hcr.xml"  mimeValue="application/hcr+xml"
description="hybrid    content    radio
recommendation list"/>
```

The receiver will find the link element, e.g. in a programme, looking for the correct MIME type, and it will trigger the enriching audio clip selection from a linked list.

Synchronisation. The enriching audio content has to be precisely synchronised with the linear content it is going to replace. The main problem with HCR is that the linear audio content could come from different sources: DAB+ digital radio, FM, or even the internet, each one with a different delay from the nominal schedule time. The simplest solution is to specify, for each transmission technology, the estimated delay between the schedule time and the actual reception time. Another proposed technique is the use of a sample-based matching technique [15] to realign the Service and Programme Information time base to the receiver's time base, independently from the transmission technology and delay.

Listeners' context and standard recommendations. The receiver has to gather accurate context (e.g. location, mood, activity) and user profile information to help create relevant content recommendations. The information can then be used to propose additional and personalised audio content to the listener. So, a standard format to describe recommendations, for example MPEG-21 user description, can be advantageous for a number of reasons. First, it enables the exchange of recommendations between trusted entities, allowing cross-domain recommendations and reducing the cold start problem. Second, it promotes an interoperable recommender engine market.

Cross-platform authentication (CPA). Nowadays, an increasing number of radio listeners switch from one device to another during the day: from a home radio, to the car radio, to the smartphone or PC. A common authentication method for all of these devices would help to maintain the listener's preferences and context across devices. The CPA [16] EBU recommendation precisely targets this need. The CPA protocol can be used to identify listeners across different devices, allowing cross-platform HCR. CPA provides a way of securely associating an internet-connected media device with an online user account. This association enables the delivery of personalised services to the device such as media recommendations, bookmarking, and pause/resume of media playback between devices. In its first version CPA focuses on devices with limited input and display capabilities such as hybrid radios (i.e. those capable of receiving broadcast radio and having a connection to the internet) but it has also been successfully applied to connected TVs and set-top boxes.

Adaptation of the enriching content. Another challenge is a proper blending between broadcast content, governed by a fixed schedule independently from the transmission media (digital radio, internet streaming), and personalised content. An optimal usage of the bandwidth is possible by properly calibrating the non-linear, personalised content to be added to the broadcast content. However, it is unlikely that additional content can be overlaid on a traditional broadcast stream without any adaptation, especially regarding the content length. The production process should include means to allow flexible content length. The object-based broadcast system [17] or similar solutions can be used during production to address the adaptation step.

The need for an interoperable radio tuner API. One key advantage of an HCR approach is that a major part of the content can be delivered with the efficient and economically sustainable broadcast channel, using the internet channel to address specific users characteristics. So, smartphones and connected radios are a natural target for an HCR service, as they often have both a broadcast radio tuner and an internet connection. However, a route to the mass market must tackle the critical obstacles of accessing the radio tuner and the lack of a standard API to access that tuner. These facts currently prevent the opportunity to take full advantage of hybrid devices for HCR. The effort to bring radio tuners and a universal API to smartphones has been led by the EBU Smart Radio initiative [18] and by the Universal Smartphone Radio Project [19].

Conclusions

The paper defined hybrid content radio as an enhanced radio service, where both linear broadcast audio and recommended audio content are used to give listeners a better radio experience, while preserving the central role of the broadcaster. HCR can both increase listener satisfaction and optimise bandwidth usage. The service prototypes created by three European broadcasters' research centres were described, exploring the potentialities of a hybrid approach for radio content. A common framework for HCR was then proposed, allowing listeners to personalise the audio of traditional linear radio with context-related audio content, addressing key requirements like metadata availability, synchronisation and authentication.

Acknowledgements

The HCR concept has been discussed in the EBU Digital Radio Platforms (DRP) group, special thanks to Lindsay Cornell, Mathias Coinchon, Michael Barroco and Ben Poor.

Bibliography

[1] KAMINSKAS M., RICCI F.: 'Contextual music information retrieval and recommendation: state of the art and challenges', *Computer Science Review*, 2012, Vol. 6, pp. 89–119

[2] Pandora. The Music Genome Project, http://www.pandora.com/about/mgp

[3] Spotify. http://www.spotify.com

[4] BURKE R.: 'Hybrid web recommender systems' in 'The Adaptive Web' (Springer-Verlag, 2007), pp. 377–408

[5] CELMA O., LAMERE P.: 'If you like radiohead, you might like this article', *AI Magazine*, 2011, Vol. 32, No. 3, pp. 57–66

[6] Ricci F., Rokach L., Shapira B., Kantor P.B. (Eds.): 'Recommender Systems Handbook' (Springer, 2011)

[7] BOZIOS T., LEKAKOS G., SKOULARIDOU V., CHORIANOPOULOS K.: 'Advanced techniques for personalized advertising in a digital TV environment: the imedia system', *Proceedings of the eBusiness and eWork Conference*, 2001

[8] CHORIANOPOULOS K., LEKAKOS G., SPINELLIS D.: 'The virtual channel model for personalised television', *Proceedings of the European Conference on Interactive Television: from Viewers to Actors?*, 2003, pp. 59–67

[9] RAUSCHENBACH U., PUTZ W., WOLF P., MIES R., STOLL G.: 'A scalable interactive TV service supporting synchronised delivery over broadcast and broadband networks', *Proceedings of the International Broadcasting Convention 2004, Amsterdam*, 2004

[10] ADOMAVICIUS G., MOBASHER B., RICCI F., TUZHILIN A.: 'Context-aware recommender systems', *AI Magazine*, 2011, Vol. 32, No. 3, pp. 67–80

[11] ANKOLEKAR A., SANDHOLM T.: 'Foxtrot: a soundtrack for where you are', *Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications*, 2011, pp. 26–31

[12] MIT BusBuzz, 2012. http://mobile.mit.edu/projects/busbuzz/

[13] MIT Loco Radio, 2013. http://www.media.mit.edu/speech/projects/locoradio/

[14] ETSI, 2015. ETSI TS 102 818, Hybrid Digital Radio (DAB, DRM, RadioDNS); XML Specification for Service and Programme Information (SPI)

[15] CASAGRANDA P., SAPINO M.L., CANDAN K.S.: 'Leveraging audio fingerprinting for audio content synchronization and replacement', *Media Synchronization Workshop*, 2015

[16] EBU, 2014. TECH 3366. The Cross Platform Authentication Protocol, Version 1.0

[17] ARMSTRONG M., BROOKS M., CHURNSIDE A., EVANS M., MELCHIOR F., SHOTTON M.: 'Object-based broadcasting – curation, responsiveness and user experience'. Proceedings of the International Broadcasting Convention 2014, Amsterdam, 2014

[18] European Broadcasting Union, Smart Radio initiative, 2014. http://www3.ebu.ch/contents/news/2014/03/radios-hybrid-future-smart-radio.html

[19] Universal Smartphone Radio Project, 2015. https://tech.ebu.ch/docs/events/radiosummit15/presentations/18_Universal%20Smartphone%20Radio%20Project%20EBU%20Digital%20Radio%20Summit%202015.pdf

[20] CASAGRANDA P., SAPINO M.L., CANDAN K.S.: 'Audio assisted group detection using smartphones'. IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2015

Selected content from IBC2015

Enhancing MPEG DASH performance via server and network assistance

Emmanuel Thomas[1] *M.O. van Deventer*[1] *Thomas Stockhammer*[2]
Ali C. Begen[3] *Jeroen Famaey*[4]

[1]*TNO, The Netherlands*
[2]*Qualcomm, Germany*
[3]*Cisco, Canada*
[4]*University of Antwerp – iMinds, Belgium*

Abstract: MPEG DASH provides formats that are suitable to stream media content over HTTP. Typically, the DASH client adaptively requests small chunks of media based on the available bandwidth and other resources. This client-pull technology has proven to be more flexible, firewall-friendly and CDN-scalable than server-push technologies. However, service providers have less control given the decentralized and client-driven nature of DASH, which introduces new challenges for them to offer a consistent and possibly higher quality of service for premium users. This issue is addressed by MPEG in a new work referred to as SAND: Server and Network-assisted DASH. The key features of SAND are asynchronous network-to-client and network-to-network communication, and the exchange of quality-related assisting information in such a way that it does not delay or interfere with the delivery of the streaming media content. MPEG is expected to publish and complete the work on SAND as a new part of the MPEG DASH standard by early 2016.

Introduction

Over the last few years, HTTP-based adaptive streaming has become the technology of choice for streaming media content over the Internet. In 2012, MPEG published a standard on Dynamic Adaptive Streaming over HTTP (DASH) [1], which has been adopted and profiled by other standards and industry bodies, including DVB, 3GPP, HbbTV and DASH-IF. The DASH formats are primarily designed to be used in client-pull based deployments with HTTP the protocol of choice for media delivery. A client first retrieves a manifest in a Media Presentation Description (MPD), and then it selects, retrieves and renders content segments based on that metadata, as seen in Figure 1.

DASH when deployed over HTTP offers some fundamental benefits over other streaming technologies. DASH requests and responses pass firewalls without any problem, like any other HTTP messages. As the content is typically hosted on plain vanilla HTTP servers and no specific media servers are necessary, DASH is highly scalable: DASH segments can be cached in HTTP caches and delivered via Content Delivery Networks (CDN), like any other HTTP content. Most importantly, a DASH client constantly measures the available bandwidth, monitors various resources and dynamically selects the next segment based on that information. If there is a reduction in bandwidth, the DASH clients selects segments of lower quality and size, such that a buffer underrun is prevented and the end user retains a continuous media consumption experience. From many studies, it is well known that start-up delays and buffer underruns are among the most severe quality issues in Internet video and DASH constitutes a solution to overcome and minimize such problems.

However, the fundamental decentralised and client-driven nature of DASH also has some drawbacks. Service providers may not necessarily have control over the client behaviour. Consequently, it may be difficult to offer a consistent or a premium quality of service. Examples include that the resources announced in the MPD may become outdated after a network failure or reconfiguration, resulting in misdirected an unsuccessful DASH segment requests by the client. A DASH client can mistakenly switch to lower quality segments, when a mobile hand-over or a cache miss is interpreted as a bandwidth reduction. Massive live DASH streaming may lead to cascades of cache misses in CDNs. A DASH client may unnecessarily start a stream with lower quality segments, and only ramp up after it has obtained bandwidth information based on a number of initial segments. Multiple DASH clients may compete for the same bandwidth, leading to unwanted mutual

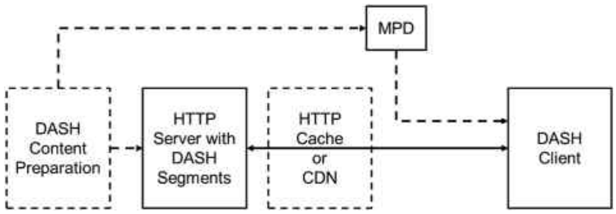


Figure 1 *Conceptual architecture of MPEG DASH.*

interactions and possibly oscillations [6]. As a consequence, service providers may not be able to guarantee a premium quality of service with DASH, even in managed networks where regular DASH clients may not fully take advantage of the offered quality of service features.

In 2013, MPEG and the IETF organised a joint workshop [2] to discuss the issues and potential solution directions, as input to the 2nd edition of the DASH standards. Soon after, MPEG started the Core Experiment on Server and Network-assisted DASH (CE-SAND), in which use cases are defined and solutions are explored. Based on the results of CE-SAND, MPEG is developing an architecture, data models and a protocol solution, expected to published as part 5 of the MPEG DASH standard in 2016. The use cases and status of the work as of mid of 2015 are summarized in the remainder of this paper.

Sand use cases and experiments

The CE-SAND addresses the following topics:

- Unidirectional, bidirectional, point-to-point and multipoint communication, with and without session management between servers/CDNs and DASH clients

- Providing content-awareness and service-awareness towards the underlying protocol stack including server and network assistance

- Various impacts on the existing Internet infrastructure such as servers, proxies, caches and CDNs

- Quality of service (QoS) and quality of experience (QoE) support for DASH-based services

- Scalability in general and specifically for logging interfaces

- Analytics and monitoring of DASH-based services

From these topics, MPEG experts derived a set of use cases to illustrate the scope of this core experiment including:

- Network mobility, e.g., when the user physically moves, which makes the device switching from one network to another

- Server failover scenario, e.g., when the segment delivery node crashes, which potentially leaves the DASH client without segments to retrieve

- Server-assisted DASH adaptation logic, e.g., when a server assists DASH clients for selecting representations

- Operational support of live service by real-time user reporting, e.g., when DASH clients report useful information in order to improve the overall quality of service

- Bi-directional hinting between servers, clients and the network, e.g., where a DASH client lets the delivery node know beforehand what it will request in the near future

- Inter-device media synchronization, e.g., when one or more DASH clients playback content in a synchronised manner

Over the past two years, MPEG experts have collected evidences from experiments and relevant feedback from the industry showing the benefits of such assistance for DASH clients. For example, one of these experiments on start-up time and quality convergence is reported in more detail below. This falls into the category of "server-assisted DASH adaptation logic".

Experiment: faster quality convergence through QoS signalling

A major source of QoE degradation in DASH is the slow convergence towards the optimal quality during the start-up period of a stream. The client will not only wait to start playback until its buffer has reached a certain size, but will also start at the lowest video quality and slowly ramp it up to find the optimal point. This will occur both at the start of a streaming session, as well as when the user switches to a different stream (e.g., channel change) or to a different point in the same stream (e.g., seeking).

A potential solution for this problem is to let the server signal the client about the performance or QoS of the connection. This enables the client to make a more elaborate choice concerning the initial quality, leading to faster convergence to the optimum. An experiment was performed to test this hypothesis. Its goals are (i) to determine the improvements in quality convergence speed as a consequence of QoS signalling, and (ii) to study any potential (adverse) side effects.

The experiment is executed using a custom built client based on libdash [3]. It implements the MSS rate adaptation algorithm described in [4], with some additional optimisations to take into account the QoS signalling variable. The `minBufferTime` is set to 8 seconds. The Big Buck Bunny video [5] of the MMSys12 dataset is used for streaming, with segment durations of

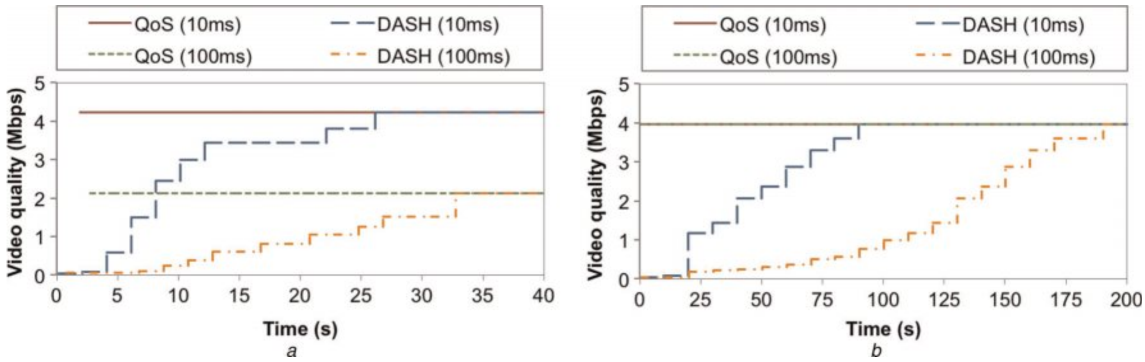


Figure 2 *Results of the QoS signalling experiment, comparing with QoS signalling and traditional DASH for 5 Mbps available bandwidth and 10 or 100 ms latency.*
a 2-second segments
b 10-second segments

either 2 or 10 seconds, and 20 quality representations between 50 Kbps and 8 Mbps. The client is directly connected to the delivery server via a traffic shaped link that has an available bandwidth of 5 Mbps and a one-way latency of either 10 or 100 ms.

Figure 2a shows the played video bitrate over time for 5 Mbps available bandwidth and 2-second segments, comparing DASH with (denoted QoS) and without (denoted DASH) QoS signalling. The figure clearly shows that QoS signalling solves the slow quality convergence process, for both low and high latencies. Traditional DASH takes, for respectively 10 and 100 ms latency, 26 and 32 seconds to reach the optimal point, whereas the DASH client with QoS signalling takes 1.8 and 1.92 seconds to start the playback at the optimal point. Playback starts a bit earlier in the traditional DASH client, however, it starts with rendering lower quality segments. If QoS signalling information is provided, the client may safely ignore the `minBufferTime*bandwidth` directive in order to keep the start-up delay comparable to rendering lower quality segments.

Figure 2b depicts the same results for 10 instead of 2-second segments. In general, the trend is similar. However, the buffer is filled up faster since there is less throughput loss due to sending fewer requests and waiting for their responses. Consequently, the initial playback delay of the approach with QoS signalling is reduced to 0.02 and 0.2 seconds for 10 and 100 ms latency, respectively. For the same reason, when using longer duration segments quality is less affected by latency in traditional DASH. Finally, the quality convergence process is more extreme without QoS signalling, clocking in at 90 and 190 seconds for 10 and 100 ms latency, respectively. As such, QoS signalling is even more advantageous when longer segments are used. Note that if there were fewer representations available to the clients, the convergence times would be shorter; however, the quality variation during the convergence period would be more drastic.

A new part for the MPEG DASH standard

The SAND architecture

In the SAND architecture, we have three broad categories of elements. They are (i) DASH clients, (ii) DASH-assisting network elements (DANE), and (iii) regular network elements. Regular network elements are DASH unaware and treat DASH delivery objects as any other object, although they could be present on the media delivery path. Transparent caches are an example of regular network elements. DASH-assisting network elements (DANE) have a minimum intelligence about DASH. For example, a DANE could identify and parse an MPD file and DASH segments to treat them differently or modify them.

The SAND architecture has the following three interfaces that carry various types of messages:

- Client-to-DANE (C2D) Interface: Metrics messages and status messages

- DANE-to-DANE (D2D) Interface: Parameters Enhancing Delivery (PED) messages

- DANE-to-Client (D2C) Interface: Parameters Enhancing Reception (PER) messages

Collectively, PED, PER, metrics and status messages are referred to as SAND messages. In this context, a media origin that serves DASH content, receives metrics messages from the clients and sends PED parameters to other DANEs is also considered a DANE element. Similarly, a third-party analytics server that receives metrics messages from the DASH clients and sends SAND messages to the clients is a DANE element. Note that the third-party server is not necessarily on the media delivery path so it does not see the DASH segments. However, as it understands the DASH metrics and produces SAND

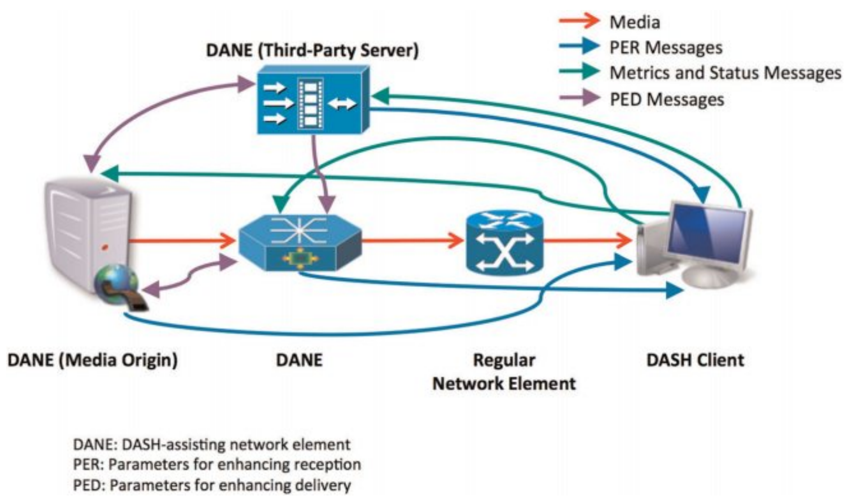

Media
PER Messages
Metrics and Status Messages
PED Messages
DANE (Third-Party Server)
DANE (Media Origin)
DANE
Regular Network Element
DASH Client
DANE: DASH-assisting network element
PER: Parameters for enhancing reception
PED: Parameters for enhancing delivery

Figure 3 *Proposed SAND architecture and message flows.*

messages for the DASH clients to improve delivery, it is still considered a DANE element.

The messages sent by the clients that carry metrics information are called metrics messages. The messages sent by the clients that carry non-metrics information are called status messages. The metrics and status messages have the same structure; however, it is important to distinguish them since these messages carry information of different nature.

The PED messages may flow in one direction or in both directions between two DANEs. This will depend on the functionality of the DANEs. For example, PED messages can be sent by a third-party analytics server to a media origin, packager or an edge router/server to enhance the delivery. However, sending PED messages the other way around (to the third-party analytics server) would not make sense as such a server only digests incoming metrics/status messages and produces PED/PER messages.

In a streaming ecosystem, there are likely several other interfaces that need to be deployed. For example, the media origin or the content provider could need an interface to the DRM servers to exchange keys. The streaming servers could need an interface to the subscription/management servers. Such interfaces are explicitly excluded from the scope of the CE-SAND.

The SAND architecture and message flows are shown in Figure 3. In this figure, metrics and status messages are shown with the same green arrows for simplicity only; this does not mean that these messages are always sent together at the same time. Note that in the figure below, the number and order of the dash-assisting and regular network elements on the media delivery path would depend on the network topology. However, this does not affect how the framework functions.

The SAND messages

While the current version of the SAND specification lists a number of PED, PER, metrics and status messages, these messages are subject to change due to the standardization process; i.e., the existing messages can be removed or modified, and new messages can be added. Yet, in this section, we provide examples from each category to demonstrate how SAND can be used in practice. These examples mainly target the use case of bi-directional hinting between servers, clients and the network.

SAND metrics and Status messages

Consider a scenario where a DASH client would like to send a hint to the cache server from which it receives media segments about which particular representation(s) of a content it is planning to fetch over a specified number of requests. For example, the DASH client may notify the cache server with a SAND Status message that it plans to request segments 41, 42 and 43 from the 5th representation for movie K. If the cache server is a DANE element that understands such a message, it can try to proactively prefetch these particular segments in advance so that they will likely be ready to be served when the actual request is received from the DASH client. This helps improve the cache hit ratio on the server as well as the streaming quality perceived by the client.

This SAND Status message, which is referred to as `anticipatedRequests`, consists of bunch of URLs (possibly with the byte-range information) for the desired segments and a target duration.

The receiving cache server may or may not be able to process these messages in time if there is an upstream bandwidth or storage shortage, or there are too many such messages all requesting different set of segments. In this case, the messages need to be prioritized appropriately and the caching algorithm may need to be adjusted for proper cache filling and eviction.

SAND PER messages

Consider a live streaming scenario where a cache is serving a large number of DASH clients. These DASH clients may have different capabilities in terms of connectivity. Prior research [6] has shown that streaming clients that share the same network resource may experience the bitrate oscillation problem, where the clients cannot figure out their fair-share bandwidth and keep requesting segments from different representations (i.e., frequent upshifts and downshifts). Later research [7] has shown that the streaming experience can quickly deteriorate even in the presence of cache servers under certain circumstances due to the lack of ability to prefetch all potential segments that could be requested by the clients.

A solution to this problem is that the cache server sends to clients a SAND PER message, referred to as `resourceStatus`, that lists what segments are available and for how long they will be available on the cache server. This provides the DASH clients a hint for their future requests to maintain a more stable streaming experience.

SAND PED messages

Consider a live sports event where the content captured in real time is encoded and packaged into a number of representations. Suppose the average representation bitrates are 1, 3 and 5 Mbps, and all the representations are available to the streaming clients. A while after the live event started, the analytics servers will start receiving metrics messages from the clients providing detailed information about their reception quality. If a large portion of the feedback indicates that most clients are oscillating between the 3 and 5 Mbps representations, the decision engine processing the analytics data can speculate that the introduction of a new representation of 4 Mbps will alleviate the problem. This decision can be conveyed to DANE-enabled transcoders and packagers through a PED message, referred to as `alterEncoding`. Alternatively, the PED message can request the transcoder to replace the 5 Mbps representation with the 4 Mbps representation, if there are constraints on processing and storage resources. Following such a change, the manifest has to be revised and the DASH clients fetch the new manifest file through usual means, possibly stimulated by a PER message.

The SAND transport protocol

In order to transport the SAND messages, an appropriate message format is necessary as well as a transport protocol. Two types of downlink scenarios are considered relevant:

1. Assistance: A scenario for which the message is provided as auxiliary information for the client, but the service will be continued even if the client ignores the message.

2. Enforcement/Error: A scenario that requires the client to act otherwise the service is interrupted. The DANE cannot or is not willing to respond to the request with a valid resource but provides suitable alternatives.

Both types of communication are relevant and it is up to the service providers to use adequate SAND messages. In addition, both DANE on the media delivery path and outside the media delivery path may use these mechanisms.

To address the use cases described in the CE-SAND, the SAND specification recommends the following HTTP-based communication channels:

- For assistance, a dedicated HTTP header field that indicates an absolute URI pointing to a SAND resource. Upon reception of an HTTP entity containing the SAND header field, the DASH client issues a GET request to the indicated element to receive the SAND message.

- For enforcement, a suitable method is the use of a 300 Multiple Choices response where the response includes an entity containing a SAND message. The entity format is specified by the media type given in the `Content-Type`.

- For error cases, a suitable method is the use of a suitable 4xx error code. The response may include a SAND message from which the client can deduce the reason for the error code and potential resolution of the problem.

The SAND communication channel can also be implemented using other protocols such as Server-Sent Events or WebSockets. In the latter, the signalling of the communication might be achieved by inserting a SAND element in the DASH MPD that advertises the URI endpoint of the communication channel as well as the corresponding protocol to use.

Conclusion

Server and Network-assisted DASH (SAND) provides a bridge between the traditional server-controlled streaming and client-controlled HTTP streaming. The technology is introduced in a way that it is expected to assist and enhance the operation of client-centric DASH streaming. The technology addresses messages as well a communication channel in order to fulfil the different requirements and use cases that were collected in the MPEG standardization process. SAND will be another step towards establishing DASH as a format that can be used for a broad set of applications and use cases.

References

[1] ISO/IEC 23009, 2012. MPEG, Dynamic Adaptive Streaming over HTTP (DASH)

[2] MPEG/IETF, 2013. *Joint Workshop on Session Management and Control for MPEG DASH*, Vösendorf, 28 July. http://mpeg.chiariglione.org/about/events/workshop-session-management-and-control-mpeg-dash

[3] Source: https://github.com/bitmovin/libdash

[4] FAMAEY J., LATRÉ S., BOUTEN N., VAN DE MEERSSCHE W., DE VLEESCHAUWER B., VAN LEEKWIJCK W., DE TURCK F.: 'On the merits of SVC-based HTTP adaptive streaming', *Proceedings of the 13th IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Ghent, Belgium, 2013

[5] Source: http://www-itec.uni-klu.ac.at/ftp/datasets/mmsys12/BigBuckBunny/

[6] AKHSHABI S., ALI C.B., DOVROLIS C.: 'An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP' *Proceedings of ACM Multimedia Systems Conf. (MMSys)*, San Jose, CA, February 2011

[7] LEE D.H., DOVROLIS C., ALI C.B.: 'Caching in HTTP adaptive streaming: friend or foe?', *Proceedings of ACM International Workshop Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, Singapore, March 2014

Selected content from IBC2015

Ready for 8K UHDTV broadcasting in Japan

Yusuke Miki   Tsuyoshi Sakiyama   Kenichiro Ichikawa   Mayumi Abe
Seiji Mitsuhashi   Masayuki Miyazaki   Akira Hanada
Kazufumi Takizawa   Ichie Masuhara   Koji Mitani
NHK, Japan

Abstract: NHK will present highly realistic broadcasts of the 2020 Olympic Games in Tokyo via 8K Super Hi-Vision, the world's most sophisticated broadcasting system. Here, we describe our preparations for 8K broadcasting, particularly for the test broadcasting via satellite due to start in 2016.

In particular, we describe our activities related to standardization and development of program production equipment, play-out, and a distribution system for the 8K broadcasting.

Regarding the production and distribution systems, we present the results of experimental satellite broadcasting performed this year. We also discuss the current status of the high dynamic range (HDR) function for the 8K system, the multi-format production workflow, and remaining issues.

Introduction

8K Super Hi-Vision (UHDTV) is a broadcasting medium featuring 16 times the number of pixels as Hi Vision (HDTV) and 22.2 multichannel sound to provide a highly realistic "you are there" sensation. Since 1995, NHK has been researching and developing UHDTV as a next-generation broadcasting system to succeed HDTV and has been active in specification studies, equipment/device development, and technology standardization. In 2012, NHK's specifications for the UHDTV video signal was approved as an international standard by the International Telecommunication Union (ITU)[1] and 8K public viewings of the London Olympic Games were successfully held in Japan, the United Kingdom, and the United States.

Now UHDTV has made a transition from the research and development stage to the implementation and promotion stage. The formulation of a roadmap toward early adoption of UHDTV (4K/8K) broadcasting has been progressing since 2012, with the Ministry of Internal Affairs and Communications (MIC) being the center of this effort. In September 2014, an interim report announced the goal of "beginning 4K/8K test broadcasting using Broadcasting Satellites (BS) in 2016 and launching 4K/8K commercial broadcasts by BS and other means by 2018 or earlier if possible."[2] In parallel with these activities, the Next Generation Television & Broadcasting Promotion Forum (NexTV-F)[3] was established, with participants including broadcasters, manufacturers of receiving equipment, and telecom companies, as an "All Japan" promotional body for UHDTV.

NHK aims to commence test broadcasting in 2016, and to this end, it is developing and preparing UHDTV facilities and equipment covering a range of functions from content production to play out, transmission, and reception. This paper describes the present state of development activities and preparations for the test broadcasting.

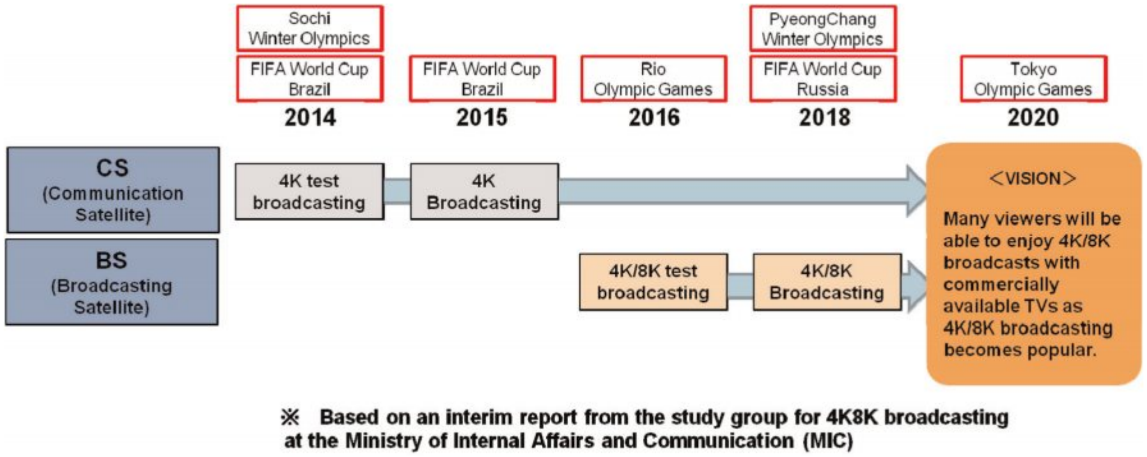
Roadmap for UHDTV broadcasting

The roadmap for 4K/8K broadcasting that appeared in "The Interim Report of the Follow-up Meeting on 4K and 8K Roadmap" of August 2014 is shown in **Figure 1**.[4] In addition to test broadcasting in 2016 and commercial broadcasting in 2018, this roadmap sets out the goals of 4K/8K broadcasting of the 2020 Tokyo Olympics and Paralympics and provision of 4K/8K programs that can be enjoyed by many viewers with commercially available TVs. NHK is making preparations in line with this roadmap.

8K UHDTV broadcasting system

Figure 2 shows the entire 8K UHDTV broadcasting system. To enable the broadcasting of live and recorded programs by 2016, NHK is preparing outside production facilities as well as editing facilities and play-out/transmitting facilities. It is also developing prototype receivers to enable viewing of test broadcasts. Testing of the technology in the field-transmission and studio facilities has already begun. The plan is to prepare these facilities so that they will eventually be capable of broadcasting.

An overview of facility preparations up to 2020 is shown in **Figure 3**. There are two major phases. The first phase, which runs up to 2016, covers development of outside production



※ Based on an interim report from the study group for 4K8K broadcasting at the Ministry of Internal Affairs and Communication (MIC)

Figure 1 *4K and 8K Roadmap*

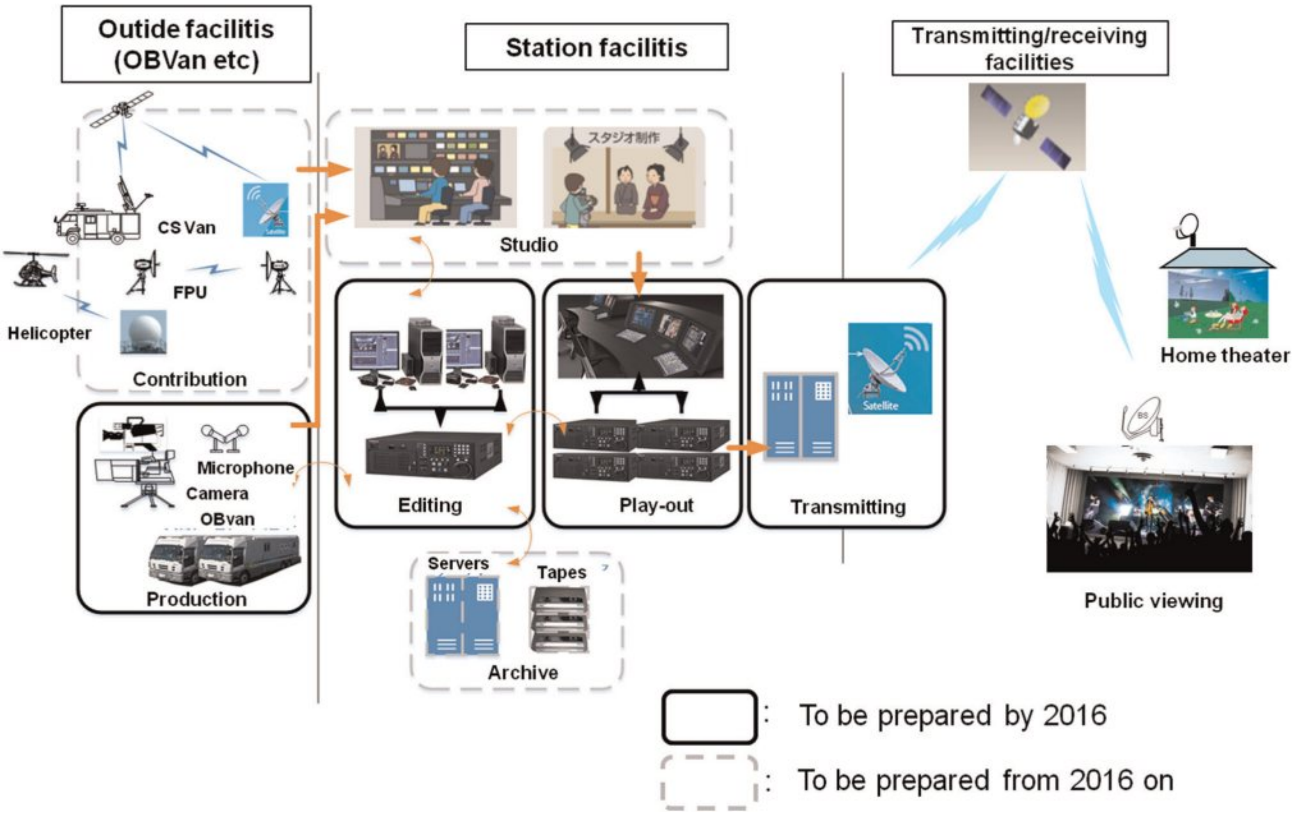


Figure 2 *UHDTV broadcasting equipment*

facilities such as cameras and outside broadcasting (OB) vans needed for content production. It also covers editing rooms, audio dubbing studios, play-out/transmitting facilities, and receiving equipment and an effort to downsize 8K equipment to make it less bulky.

The second phase, from 2016 on, will expand and enhance facilities and improve the system performance and reliability toward broadcasting in 2018 and the Tokyo Olympics and Paralympics in 2020. It also covers improvements to equipment needed for supporting a wide colour gamut (ITU-R BT.2020)[5], high dynamic range (HDR), high frame rate (HFR), etc. The goal for the facility preparations is to lower the cost of equipment by making use of 4K technology where possible and unifying facility specifications.

Preparations policies

The main policies governing the preparation of UHDTV facilities are as follows.

- Steady development and preparation of play-out/transmitting facilities and receiving equipment toward the launch of test broadcasting in 2016

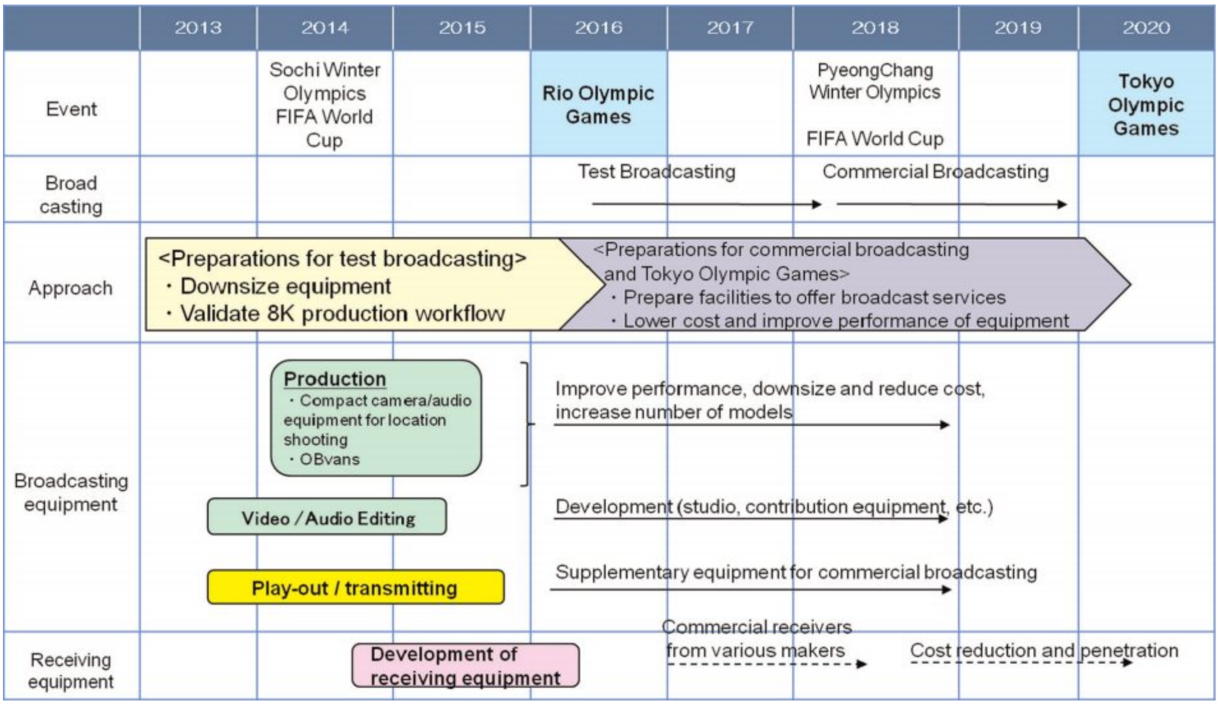

2013
2014
2015
2016
2017
2018
2019
2020
Event
Sochi Winter Olympics FIFA World Cup
Rio Olympic Games
PyeongChang Winter Olympics
FIFA World Cup
Tokyo Olympic Games
Broad casting
Test Broadcasting
Commercial Broadcasting
Approach
<Preparations for test broadcasting>
・Downsize equipment
・Validate 8K production workflow
<Preparations for commercial broadcasting and Tokyo Olympic Games>
・Prepare facilities to offer broadcast services
・Lower cost and improve performance of equipment
Broadcasting equipment
Production
・Compact camera/audio equipment for location shooting
・OBvans
Improve performance, downsize and reduce cost, increase number of models
Video / Audio Editing
Development (studio, contribution equipment, etc.)
Play-out / transmitting
Supplementary equipment for commercial broadcasting
Receiving equipment
Development of receiving equipment
Commercial receivers from various makers
Cost reduction and penetration

Figure 3 *Roadmap for preparing Super Hi-Vision equipment*

- Provision of highly mobile production equipment enabling diverse program production

- Improvement of facility functionality and performance for efficient content production and shorter production times

- Enhancement of outside production facilities such as video/audio OB vans

At NHK, we are concentrating on developing play-out/transmitting facilities for test broadcasting in the next year. Furthermore, in anticipation of broadcasting in 2018 and broadcasts, public viewings, and overseas distribution of programming of the Tokyo Olympics and Paralympics, we plan to construct reliable multifunctional facilities of the system in a stepwise manner.

Continuous efforts must be made to raise the performance and functionality of production facilities to enable the production of diverse 8K content. To date, 8K cameras and recorder equipment have been bulky and severely deficient in their mobility as tradeoffs for their abilities to handle high-data-rate video/audio signals. We are therefore attaching great importance to improving mobility, downsizing, and reducing the power consumption of production equipment for practical use.

Additionally, for postproduction needs, we are preparing video editing and audio dubbing studios to create an environment in which high-quality content can be produced in as short a time as possible. In the video editing room, offline data consisting of video material converted to HD will be used as an efficient means of inputting 8K material into editing equipment. The audio dubbing studios will support not only 22.2 multichannel sound but also stereo and 5.1 multichannel productions.

Live broadcasts of sports events, music concerts, etc. are the sort of programming that can take full advantage of the features of 8K UHDTV. To expand and enhance content production functions for such outside broadcasting, we are developing video OB vans that can carry a maximum of ten cameras. We have already developed an audio OB van for production of 22.2 multichannel sound within the vehicle.

The following introduces preparations that we are making in line with the above policies and major facilities now in development.

Play-out/transmitting facilities and receiving equipment

We are developing UHDTV broadcasting facilities in keeping with the “Transmission System for Advanced Wide Band Digital Satellite Broadcasting” standard[6] developed by the Association of Radio Industries and Businesses (ARIB) in Japan and the operating provisions presently being formulated at NexTV-F. An overview of play-out/transmitting facilities for test broadcasting is shown in **Figure 4 and Figure 5**.

Using a Broadcasting Satellite (BS), these facilities can transmit a single 8K program or two 4K programs in a multi-level format (main-channel/sub-channel) simultaneously with 22.2ch, 5.1ch, and 2ch audio channels (32 audio channels maximum). Here, we will construct core

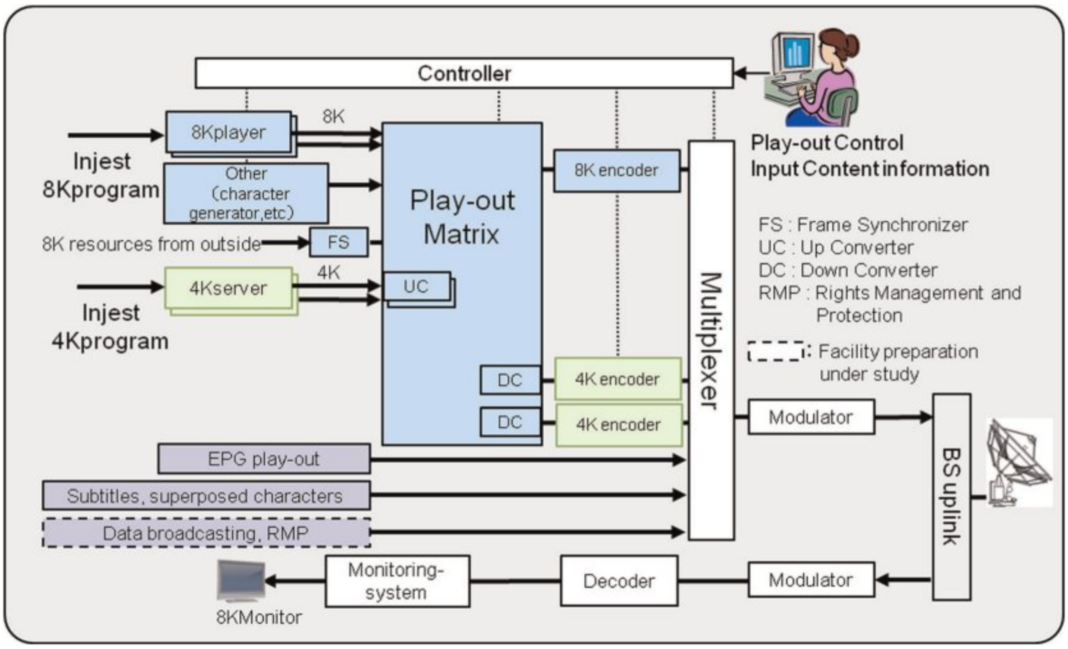
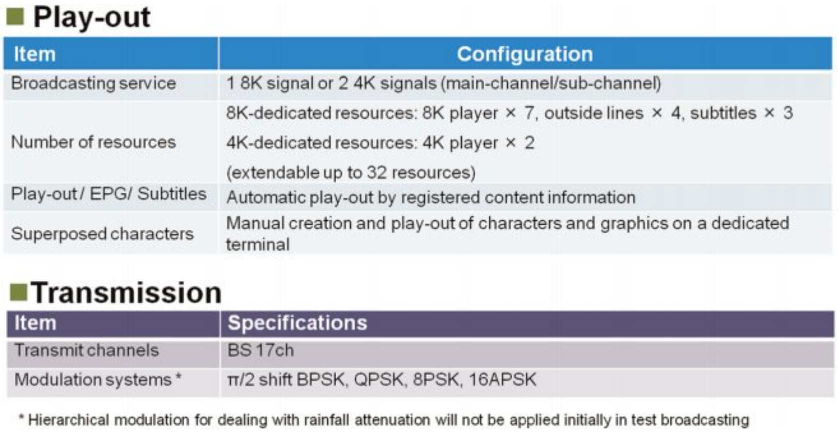


Figure 4 *Overview of Super Hi-Vision play-out/transmitting equipment for test broadcasting*



■ Play-out

Item	Configuration
Broadcasting service	1 8K signal or 2 4K signals (main-channel/sub-channel)
Number of resources	8K-dedicated resources: 8K player × 7, outside lines × 4, subtitles × 3 4K-dedicated resources: 4K player × 2 (extendable up to 32 resources)
Play-out / EPG/ Subtitles	Automatic play-out by registered content information
Superposed characters	Manual creation and play-out of characters and graphics on a dedicated terminal

■Transmission

Item	Specifications
Transmit channels	BS 17ch
Modulation systems *	π/2 shift BPSK, QPSK, 8PSK, 16APSK

* Hierarchical modulation for dealing with rainfall attenuation will not be applied initially in test broadcasting

Figure 5 *Overview of Super Hi-Vision play-out/transmitting facilities specifications*

facilities for 8K equipment that will up-convert (UC) and down-convert (DC) input/output 4K signals so that all video signals can be routed in a uniform manner as 8K signals.

We are also preparing a variety of multimedia services in addition to video and audio services. We have already begun development of an electronic program guide (EPG) and captioning/subtitling functions, and we plan to implement a data broadcasting service, a copy-protection function, and other services over time.

To efficiently transmit high-capacity and high-quality video/audio information in UHDTV broadcasts, we are adopting new source coding schemes such as High Efficiency Video Coding (HEVC) and MPEG4 Advanced Audio Coding (AAC). We are also looking to introduce advanced technologies such as the newly developed MPEG Media Transport (MMT) multiplexing system, the latest browser standard (HTML5), and other facilities that can enable diverse broadcasting services tailored to the digital age.

In parallel with the above efforts, we are also developing receiving equipment so that public viewings can be held at NHK broadcasting stations throughout Japan and as many people as possible can enjoy test broadcasts beginning in 2016. The development of thin, compact 8K monitors has been accelerating in recent years, and we are working to make the receiver section of these monitors as compact as possible by introducing advanced technologies (**Figure 6**).

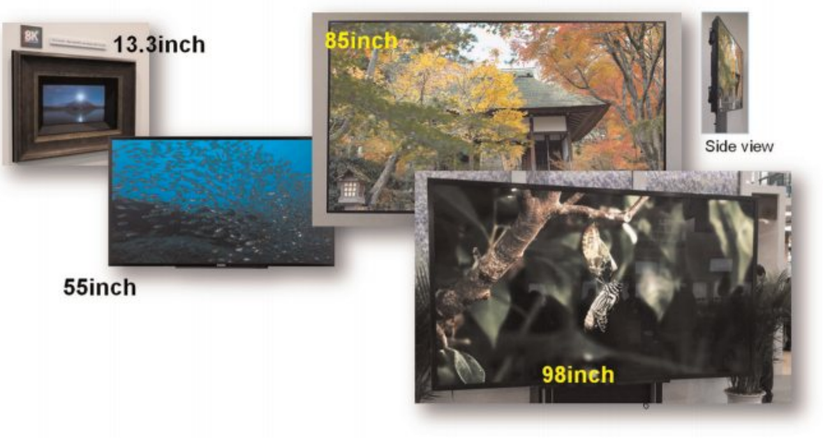


Figure 6 *8K Monitors*

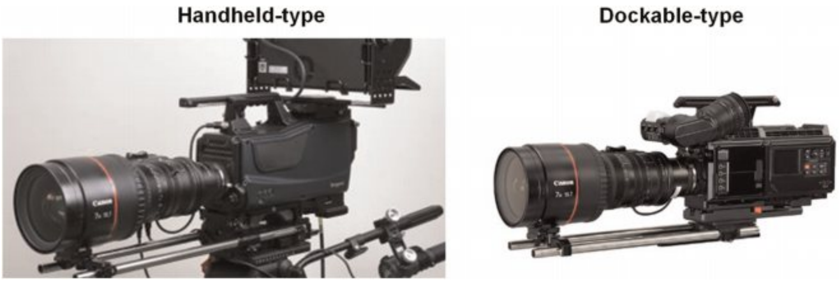


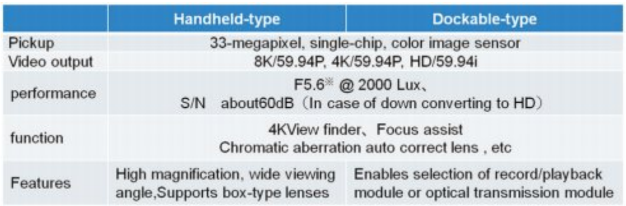
Figure 7 *Appearance of compact cameras*

Production facilities

In the commercialization stage, it is important that the equipment have high operability reflecting experience of actual production in addition to high functionality and performance. The appearance and specifications of new 8K cameras developed in 2015 are shown in **Figures 7** and **8**, respectively. The specifications of these cameras reflect improvements on the previous cameras and the results of pickup tests conducted with prototype cameras at production sites.

The handheld camera features a built-in 4K viewfinder and functional improvements such as automatic lens chromatic aberration correction. It is also supersensitive and quiet (reduced fan noise) for shooting in theaters. The dockable-type camera, meanwhile, enables a wide range of shooting styles thanks to it having a separate camera head section (weight: 3 kg) and function-extension unit attached to the rear of the head. Changing the function-extension unit enables the camera to be used as a camcorder, a relay camera, and other functions.

We have also developed an UHDTV Recorder, the appearance and specifications of which are shown in **Figures 9** and **10**. While the previous model needed 16 memory cards for recording, the new model needs only four memory cards, thanks to its high compression efficiency and high-speed, large-capacity memory. The main unit is about one-third the size of the previous unit. We also developed equipment for making backups from memory to tape media using Linear Tape-Open (LTO) technology. Furthermore, while the backup time from memory to tape media was about eight times the recording



	Handheld-type	Dockable-type
Pickup	33-megapixel, single-chip, color image sensor	
Video output	8K/59.94P, 4K/59.94P, HD/59.94i	
performance	F5.6※ @ 2000 Lux、 S/N about60dB（In case of down converting to HD）	
function	4KView finder、Focus assist Chromatic aberration auto correct lens , etc	
Features	High magnification, wide viewing angle,Supports box-type lenses	Enables selection of record/playback module or optical transmission module

Figure 8 *Basic specifications of compact cameras*

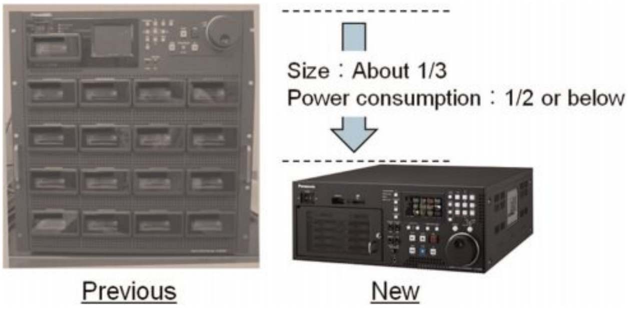
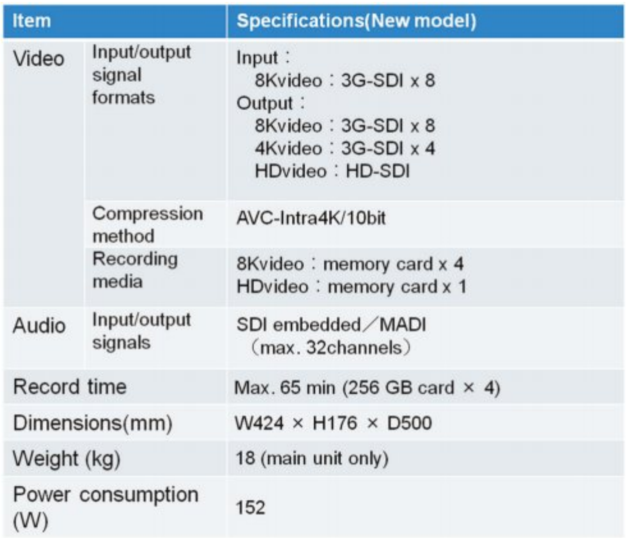


Figure 9 *Appearance of UHDTV Recorder*



Item		Specifications(New model)
Video	Input/output signal formats	Input： 8Kvideo：3G-SDI x 8 Output： 8Kvideo：3G-SDI x 8 4Kvideo：3G-SDI x 4 HDvideo：HD-SDI
	Compression method	AVC-Intra4K/10bit
	Recording media	8Kvideo：memory card x 4 HDvideo：memory card x 1
Audio	Input/output signals	SDI embedded／MADI （max. 32channels）
Record time		Max. 65 min (256 GB card × 4)
Dimensions(mm)		W424 × H176 × D500
Weight (kg)		18 (main unit only)
Power consumption (W)		152

Figure 10 *Basic specifications of UHDTV Recorder*

time with the previous equipment, the new equipment shortens it to about three times.

In addition, we have been developing equipment to enable conversion between 8K and 4K/HD video signals so that our production system will be as efficient as possible.[7]

OB van

We developed two OB van models each equipped with up to ten cameras to ensure the same scale of relay performance as that of existing HD OB vans. The vans incorporate 8K switchers with 16 or more inputs to make it easy to increase the number of in-vehicle cameras and special-effects equipment such as for slow-motion playback. The appearance and specifications of the OB vans are shown in **Figures 11** and **12**. The production room also has an extension function on one side to provide enough space for an operator of slow-motion playback equipment, for example.

In 2015, we also developed an audio OB van for 22.2-multichannel sound production. The vehicle is 11.5 m long, 2.5 m wide, and 3.5 m high. By widening the van and lengthening the workspace, we created a 22.2-multichannel mixing room in which speakers are arranged on a 2.1m-radius spherical surface centred on the mixer (**Figure 13**). The van can handle 5.1ch and stereo sound.

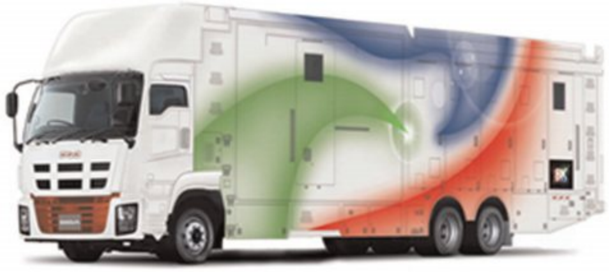
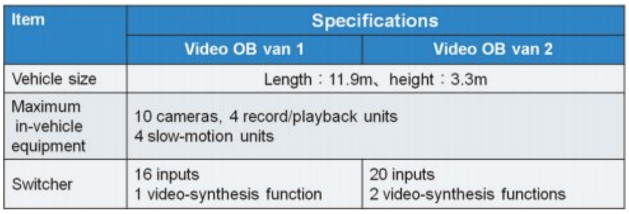


Figure 11 *Appearance of OB van*



Item	Specifications	
	Video OB van 1	Video OB van 2
Vehicle size	Length：11.9m、height：3.3m	
Maximum in-vehicle equipment	10 cameras, 4 record/playback units 4 slow-motion units	
Switcher	16 inputs 1 video-synthesis function	20 inputs 2 video-synthesis functions

Figure 12 *Basic specifications of an OB van*

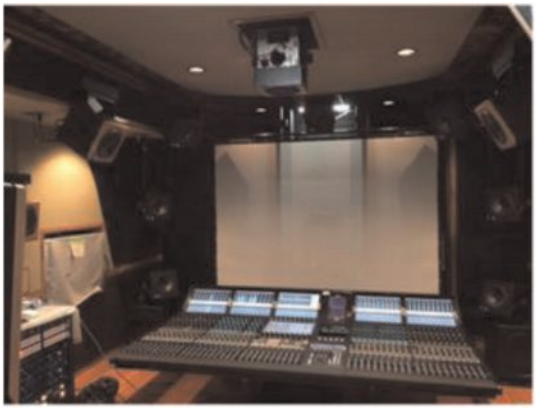


Figure 13 *Mixing room in audio OB van*

Content production and other industrial use

NHK has a proven track record in public viewings (**Figure 14**) and other presentations of 8K live broadcasts, starting with the London Olympic Games in 2012 and continuing with the Sochi Winter Olympics, FIFA World Cup in Brazil in 2014, the FIFA Women's World Cup in Canada and Wimbledon Tennis in 2015. We have also accumulated diverse 8K content through on-location shoots of nature and travel programs, recording entertainment presented in theatres and concert halls. Moreover, we are studying how 8K can be used in a variety of areas besides broadcasting, including in medicine (8K-video recording of

Figure 14 *8K Public viewing*

heart surgeries and 8K endoscopes) and in education (electronic blackboards), in digital signage and so on.

Future facilities preparations and remaining issues

In anticipation of the 2020 Tokyo Olympics and Paralympics, NHK will work to expand and enhance its facilities.

The following issues must be addressed with an eye toward widespread use of UHDTV broadcasting:

- Development and preparation of wireless transmission facilities, studio facilities, etc.

- Further performance gains in the production system

- Commoditization and price reduction of equipment

The type of 8K broadcasting services to provide through an all-Japan system consisting of broadcasters, manufacturers of receiving equipment, etc., is now a subject of discussion. Finding an answer to this question will require studies on the scale of facilities and the development of new equipment and systems in accordance with service requirements. These developments may include an 8K satellite OB van (8K-CSvan and 8K microwave link(8K field pickup unit (8K-FPU)) to simplify relays of sports events and news stories and the transfer of video materials as well as an 8K studio for switching between outside and in-station resources in program production. Preparation of such facilities will require studies on specifications suitable for 8K production, such as transmission schemes and frequencies that can transfer large amounts of data without delay, camera arrangements and studio size/designs conducive to 8K studio productions, etc.

How to extract further performance gains for the production system is the major issue here. The present system has a signal format that achieves 8K video through four 4K image sensors each corresponding to a different color channel (GGRB). This system can transmit 8K signals efficiently while preserving resolution in the horizontal and vertical directions, but it suffers from degraded resolution in the diagonal direction. From 2016 on, we plan to use an 8K signal format capable of even higher picture quality (8K/YCbCr 4:2:2). We are also studying new functions such as high dynamic range (HDR) support and will strive to implement such functions as early as possible. In a parallel way, We examine how to product programs in the situation of mixed HD/4K/8K equipment.

Finally, the commoditization and price reduction of systems and equipment will be essential if 8K UHDTV is to become popular. The commercialisation of Hi-Vision is potentially illustrative; reception equipment that initially cost several hundred thousand dollars (at the current exchange rate) eventually cost about 40,000 dollars by the time of test broadcasting in 1991[8], which despite being costly was low enough for popularization to take hold. Similarly, for 8K UHDTV, we expect the number of equipment shipments to increase at the start of test broadcasting and prices to drop as a result. We also plan to continue developing equipment and devices with cost in mind while making use of existing HD, 4K, and information-communications technology where possible to bring costs down even further.

Conclusion

This paper described the state of development of 8K UHDTV facilities with a focus on facilities for test broadcasting scheduled to begin in 2016. At NHK, we have undertaken the development of facilities to facilitate diverse program production and have been working to downsize and improve equipment performance such as new and advanced compact cameras, UHDTV recorder, etc. Going forward, our plan is to promote the spread and adoption of equipment and devices that will enable more producers to become involved in the production of 8K programs and content. In addition, the development of play-out/transmitting facilities is progressing in parallel with the formulation of standards and operating provisions. Over the next 12 months, we can expect a massive number of integration tests for separately developed equipment and system trials. We must achieve the world's first 8K broadcasting system. Moreover, in conjunction with test broadcasting to begin in 2016, we would like to construct an environment that will enable many people to enjoy 8K UHDTV on prototype receivers installed at NHK broadcasting stations throughout Japan.

By continuing its leading role in promoting 8K UHDTV and working to make equipment and facilities ready for practical use, NHK is committed to delivering highly realistic, high-presence content of the 2020 Tokyo Olympics and Paralympics to all its viewers through a technologically advanced broadcasting system.

References

[1] Rec. ITU-R BT.2020, 2012. Parameter Values for Ultra-High Definition Television Systems for Production and International Programme Exchange

[2] Ministry of Internal Affairs and Communications (MIC): 'The interim report of the follow-up meeting on 4K and 8K roadmap', September 2014 (in Japanese)

[3] Next Generation Television & Broadcasting Promotion Forum (NexTV-F). http://www.nextv-f.jp/

[4] Ministry of Internal Affairs and Communications (MIC): 'Follow-up meeting on 4K and 8K roadmap (5th meeting)', Handouts (in Japanese). http://www.soumu.go.

jp/main_sosiki/kenkyu/4k8kroadmap/02ryutsu11_03000039.html

[5] Masaoka 'Color management for wide-color-gamut UHDTV production', *SMPTE 2014 Annual Technical Conference & Exhibition*

[6] ARIB STANDARD ARIB STD-B44: 'Transmission System for Advanced Wide Band Digital Satellite Broadcasting'

[7] Ichikawa *et al.*: 'Development of UHDTV (8K) baseband processor unit "BPU-8000"', *SMPTE 2014 Annual Technical Conference & Exhibition*

[8] Ministry of Posts and Telecommunications: 'Information and Communications in Japan White Paper, 1991 Edition', Japan's Social and Economic Life and Information and Communications Technology, Chapter 2, Section 4 (in Japanese)

Selected content from IBC2015

The impact of subtitle display rate on enjoyment under normal television viewing conditions

J. Sandford

The British Broadcasting Corporation, UK

Abstract: One of the properties often identified as having an impact on the television viewing experience for subtitle users is the rate of subtitling (measured in words per minute) (1,2). Previous studies on the subject have often restricted participants from using residual hearing or lip-reading as they naturally would when viewing television (3,4,5,6). Additionally, some studies were carried out with potentially biased participants (5,6). No research has been done to date at a large scale on the rate of scrolling subtitles as are often used in live subtitling (5,6).

This paper presents the results of a study examining the impact of subtitle display rate on enjoyment for a representative sample of subtitle users. Specially created and off-air material was used with both block and scrolling subtitles. Sound was available and lip-reading was possible. The results challenge previous assumptions.

Introduction

The rate of subtitles is often highlighted in subtitling guidelines as an important factor in viewer understanding and enjoyment (2), but no scientific justification is provided. When the few papers currently available on the subject of subtitle rate are examined, it becomes apparent that the quality of previous research is poor and findings vary wildly as a result. Furthermore, research in the field repeatedly fails to use un-biased regular subtitle users (i.e. people who use subtitles once a day or more) as participants and fails to use normal television viewing conditions (3,4,5,6).

This paper presents the findings of a new study that improves on previous work and answers some questions while querying the validity of others.

- What is the ideal rate of subtitles for subtitle users?
- At what subtitle rates is enjoyment diminished for subtitle users?
- How do these rates compare to the enjoyment of speech at various rates for hearing viewers?
- Does the rate of subtitles even have an impact on enjoyment?

Background

Subtitle rate measurement

Subtitle rate (also known as the speed of subtitles) is most often measured in Words Per Minute (WPM). This may be calculated in a number of ways. The most common method used is to take an average over a period of time by dividing the number of words in a clip or programme by its length in minutes. This method is used in much of the available academic literature. While this is a simple method to implement, it may provide low readings for clips with long periods without speech. This may be accounted for by excluding long periods of silence from calculations. Studies generally choose their clips carefully to avoid this problem.

The measurement of rate in this study used this method. Clips in part 1 of the study had no periods without subtitles. Periods without subtitles in the clips in part 2 were excluded from calculations.

Subtitle rate in guidelines

Guidelines often quote optimal and maximum rates for subtitles. Figures of approximately 140WPM as the optimum subtitle rate and around 180-200WPM as the maximum rate are common. The guidelines examined fail to cite research supporting these figures but justify them by stating that above these rates, subtitles will be difficult to follow (2).

Prior research

The small amount of published research on subtitle rate varies wildly in quality. Participants are sometimes selected from biased or non-representative groups. These include people who work in subtitling, people who do not use subtitles and people from interest groups who may subconsciously aim to represent the standard views of their group (5,6,7). Many studies also purposefully aim to reduce experimental influences to the subtitles alone by using footage without those speaking in shot or by using clips without audio (3,4,5,6). This has the un-desirable side effect of creating an un-natural viewing experience. Viewers normally use visual cues such as lip-reading or facial expression to support subtitles. Most subtitle users also have some level of hearing and thus use subtitles in conjunction with audio. Viewers' experience is a combination of these sources of information.

Previous research has shown no drop in comprehension at rates of at least 230WPM (3,4,7), far higher than the maximum rates in current guidelines. One study which aimed to find the most enjoyable rate of subtitles identified a speed of 145WPM, which is approximately the average speed of American subtitles found in a study conducted by the same researcher (3,8). However, the materials in this study used footage without people speaking in shot and without audio.

Requirements for this study

In order to identify maximum, minimum and optimum subtitle rates, this study built upon the method of Jensema 1998 (3). Participants were presented with clips of a range of speeds and asked to rate the speed and their enjoyment of the subtitles. In addition, a control group of hearing viewers were asked to rate the speed and their enjoyment of the speech on the same clips but without subtitles. Where Jensema used un-natural clips, this study replicates normal television viewing conditions with the speakers in shot and audio available. The two main display methods of subtitles, block and scrolling (also known as word-at-a-time), were tested. To identify if these results held for real-world content, a range of off-air (broadcasted) clips identified as being far faster than current guidelines were also tested.

Study – Part 1

For Part 1 of this study, clips were created at a range of rates with both block and scrolling subtitles. Subtitle users were asked to rate these in terms of speed and enjoyment. This would allow the optimum rate of subtitles and rates at which subtitles become too fast or too slow to be identified. The same rates would also be identified for speech for hearing viewers.

Materials

24 news clips were created for this study. A local news team, studio and presenter were used to keep the style consistent, familiar and realistic. Both recorded audio and video were used in the clips with no other audio or visual content added.

All clips were 30 seconds long to eliminate differences in fatigue between clips. The number of words was therefore changed to alter the rate (e.g. a 170WPM clip would be scripted to 85 words). Speech and subtitles were correctly aligned in all clips to eliminate the confusing effects introduced by differing speech and text for those who used lip-reading or audio in conjunction with the subtitles. Rates of 90, 110, 130, 150, 170, 190, 210 and 230WPM were used which approximate those used in Jensema 1998 (3). Three sets of 8 clips were created with one set shown with block subtitles, one with scrolling and one used for introductory materials for each participant. Having 3 clips at each rate would also reduce any effects of individual scripts. Realistic scripts were created by identifying local news stories consisting of approximately the number of words required and rewording sections to make the length correct. The stories were purposely chosen from regions other than the one the study was conducted in to reduce the possibility of familiarity with the stories. A 3 second still of a black background with a dark logo was displayed before and after each clip to allow the participant to comfortably switch between rating and viewing clips.

Where subtitles were used, their style and layout were matched to that of BBC News. Splitting of lines was based on the maximum number of characters allowed in a subtitle and not on grammatical boundaries to avoid effects of artistic choices and matched news subtitling styles. All clips were subtitled verbatim and had versions with both block and scrolling subtitles produced. In the case of scrolling subtitles, each word was introduced on its first spoken utterance. In the case of block subtitles, each subtitle was introduced on the first spoken utterance of the first word. The final subtitle of each clip was removed at the point the final utterance finished. There were no breaks in the subtitles. The speed of each clip used in analysis was the measured rate, not the target rate. Where clips were shown without subtitles, the subtitled rate and not the spoken rate (which may be higher for numbers etc.) was used in analysis to allow direct comparisons to be made with the subtitled clips.

Methodology

The study was conducted with the participant seated at a distance of approximately 5H from the television (where H is the height of the television) (9). A table was provided with a mouse to allow the participant to interact with the user interface for questions presented on the television. The television remote was also provided for setting the volume.

For each participant the three sets of videos were assigned as introductory material, block subtitles and scrolling subtitles. Additionally, half of the participants were shown their block subtitles set first and half their scrolling

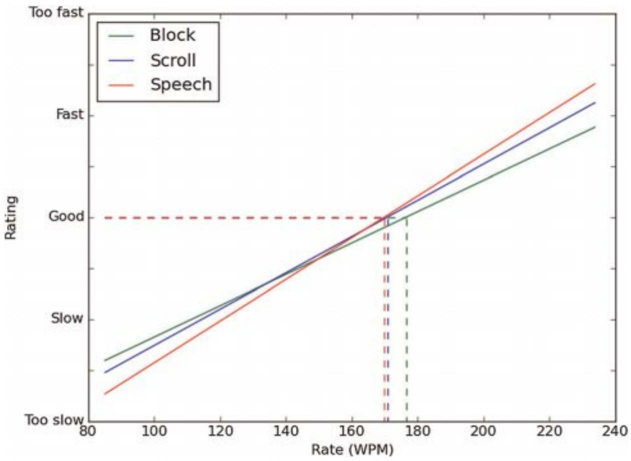


Figure 1 *Perceived rate of subtitles and speech against measured rate*

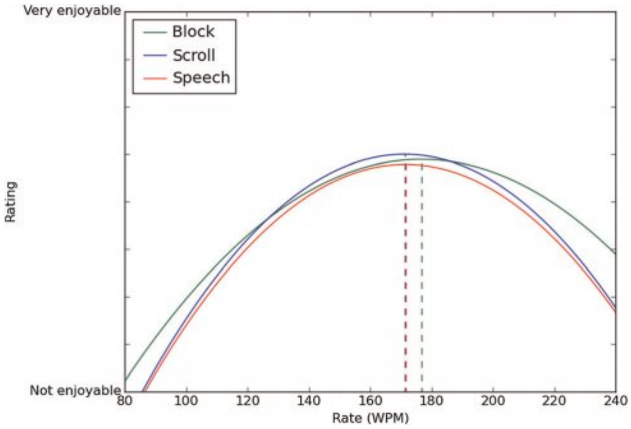


Figure 2 *Perceived enjoyment of subtitles and speech against measured rate*

subtitles set first. Participants were first shown the 150WPM clip in their introductory set with block subtitles and asked to set the volume of the television as they would normally have it when watching television with subtitles. For each main set, they were then shown 3 introductory videos of slow (90 or 110WPM), medium (170 or 190WPM) and fast (210 or 230WPM) rates in that order with the subtitling format (scrolling or block) of that set; the lower speed was shown with the first set. Participants were asked to rate each clip in terms of speed ("Too slow" to "Too fast") and enjoyment ("Not enjoyable" to "Very enjoyable") of the subtitles on continuous scales with labels at each end only. Each question was displayed separately to reduce cross-rating interference. The clips in the main set were then shown, ordered according to a Latin-square. Participants were asked to rate these clips as before.

A control group of hearing participants were asked to set the volume and shown introductory content as with the subtitles group but were only asked to rate one other set of clips, not two. Any wording on screen that referred to subtitles in the main group referred to speech in the control.

Participants

25 frequent subtitle users were recruited through an external agency. A split of male and female participants were recruited along with a spread of ages, hearing impairments and social grades. All were regular users of subtitles as an access service and were familiar with televised BBC News content. No participants who use sign language as a first language were recruited. First language BSL users, who make up around 8% of hearing impaired people in the UK (10), may be seen as second language users and will require a specific detailed study. A convenience sample of 16 hearing participants not involved in production quality were recruited from BBC North at MediaCityUK for the control group.

Results

Figures 1 & 2 show the ratings for speed and enjoyment for both groups of participants. Table 1 shows exact values where mean rate is on a scale where 1 is "Too slow" and 5 is "Too fast". "Slow", "Good" and "Fast" rates are taken from a linear regression at ratings of 2, 3 and 4 respectively. Mean enjoyment is on a scale where 1 is "Not enjoyable" and 5 is "Very enjoyable". Peak enjoyment was found with a quadratic regression.

The optimum ("Good") rate was found to be highest for block subtitles and was approximately the same for scrolling subtitles and speech. The range of rates between "Fast" and "Slow" was widest for block subtitles and narrowest for speech. However, the overall similarity between all of these results demonstrates that the rate of subtitles is not an issue under the conditions tested. When the rate of speech is perceived to be bad, the rate of subtitles is also perceived to be bad. Also, when the rate of speech is perceived to be good, the rate of subtitles is perceived to be good.

Table 1 – Study Part 1 results for perceived rate and enjoyment

	Rate R^2	Rate mean (1-5)	Rate "Slow" (WPM)	Rate "Good" (WPM)	Rate "Fast" (WPM)	Enjoy R^2	Enjoy mean (1-5)	Enjoy peak (WPM)
Block	0.64	2.77	112	177	242	0.35	2.86	177
Scrolling	0.68	2.84	115	171	227	0.42	2.76	171
Speech	0.77	2.83	121	170	219	0.39	2.67	171

Study – Part 2

Part 1 of the study has shown that, under the conditions tested, the rate of subtitles is not an issue. Part 2 aims to explore if this remains true for broadcast content. A range of clips from broadcast content identified as having subtitle rates above 200WPM were selected to see how their ratings compare to the material in Part 1.

Materials

8 clips identified using a monitoring system were selected to cover a range of programming styles. These contained differing numbers of people talking and varying shots such as close-ups, long shots and shots of people/content other than the person talking. The clip lengths, styles, and mean instantaneous subtitle rates for each clip are shown in Table 2.

The subtitles for these clips were shown as they were presented when broadcast. All timing, styling and positioning was maintained including any inaccuracies. All clips had pre-prepared block subtitles. No live scrolling subtitles were used due to the nature of how live subtitles are created. They are often created by a single subtitler speaking the subtitle content into speech recognition software and carrying out minor formatting with a keyboard. This method means live subtitles rarely reach the highest subtitling rates.

Methodology

These clips were shown to the regular subtitle users immediately after part 1. The same rating system was used and the clips were shuffled using a Latin-square.

Results

Figures 3 & 4 show the ratings of part 2 against the regression lines calculated in part 1. Exact values are shown in Table 2. All mean perceived rates are closer to “Good” than the prediction from part 1 and well under the “fast” mark. The mean perceived rate of 3 clips subtitled at approximately 230WPM fall within 1% of a perceived perfect rate. Mean enjoyment is also well above the prediction from part 1 for all clips.

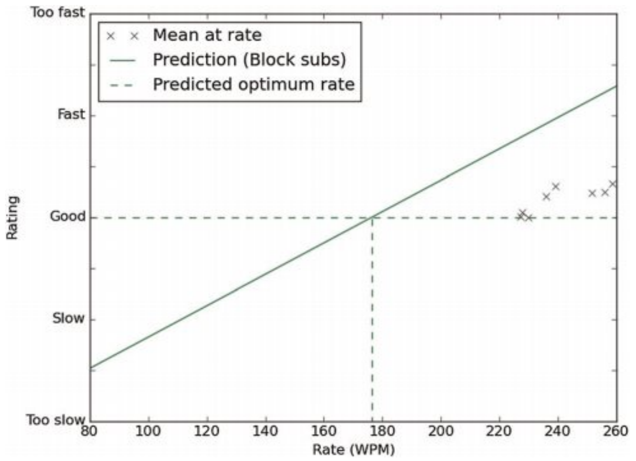


Figure 3 *Perceived rate of off-air content against measured rate*

These results show that the perceived ideal rate found in part 1 does not apply to all content. Different content felt right at different speeds. Any difference in perceived rate must therefore be related to other issues. Table 2 also shows the spread of the data for each clip in the form of the inter-quartile range (IQR). Some clips have far larger ranges than others. This may indicate that personal preference or personal resilience to other issues within subtitles has a large effect on perceived rate and enjoyment.

It should be noted that a technical error resulted in only 21 of the 25 participants viewing the weather clip.

Discussion

Part 1 of this study found the optimum rate of subtitles to be 171WPM for scrolling subtitles and 177WPM for block subtitles. These rates are approximately the maximum

Table 2 – Study Part 2 clip information, and means and IQRs for perceived rate and enjoyment

Title	Style	Clip length	Rate (WPM)	Rate mean (1-5)	Rate IQR (1-5)	Enjoy mean (1-5)	Enjoy IQR (1-5)
Rockford	Drama	3m06s	252	3.24	0.28	3.51	1.41
Homes	Factual	0m56s	227	3.01	0.05	4.00	1.42
Topgear	Talk show	0m56s	256	3.24	0.49	3.40	1.54
Weather	News	0m59s	239	3.31	0.38	3.01	0.61
Wogan	Talk show	2m04s	230	2.99	0.06	3.85	1.32
Escape	Factual	2m11s	228	3.05	0.05	3.84	1.63
Kitchen	Cookery	1m04s	259	3.33	0.63	2.34	1.17
Perfection	Quiz	0m56s	236	3.21	0.23	3.40	0.78

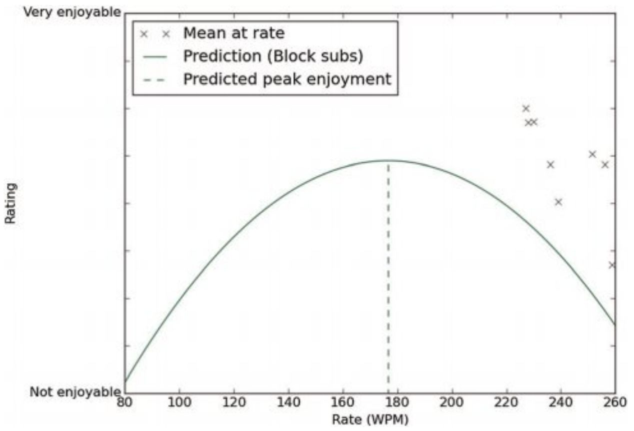


Figure 4 *Perceived enjoyment of off-air subtitles against measured rate*

currently allowed under some guidelines. The point at which most subtitle users would find the subtitles too fast was found to be 227WPM for scrolling and 242WPM for block subtitles. These values are far above the 200WPM some guidelines warn would be difficult for many viewers to follow (2). Furthermore, Part 1 of this study went on to find that ratings for subtitles at various rates were close to or better than those for hearing viewers rating the speed of speech on clips without subtitles. From this we infer that when there is an issue with the rate of subtitles, the same issue can be expected in speech and vice-versa. Issues of rate may then be expected to be noticed by content producers in the speech before the subtitles are created. Further to this, participants within this study repeatedly commented that "if there is a mismatch [between the subtitles and speech] then that's a problem". They were confused by our request for them to rate the speed of subtitles as this is dictated by the speed of speech and cannot be changed. Many see it as important that the subtitles are as close as possible to the speech in both timing and wording to make the understanding of speech/lip-reading and the subtitles combined as easy as possible. Part 1 of this study showed that not only is the rate of subtitles not an issue when the subtitles are verbatim and correctly timed, but that the way that the issue of rate is interpreted by previous academic literature and guidelines does not match the perceptions of users.

Part 2 of this study aimed to discover if the findings in part 1 held true across a range of real-world content. Clips as high as 230WPM in this section were tightly rated as perfect in speed. The three best rated clips contained discussions between multiple people, a situation where the overall rate of speech is naturally higher. Furthermore, some of these clips occasionally had the speaker out of shot or in wide shots. The consistently good ratings suggest that people's following of the content was not impaired greatly by the inability to lip read for short periods. That said, some participants did express a preference for the speaker's face to be in shot. This not only enables lip-reading but also the interpretation of emotions absent from the subtitles. This section of the study clearly demonstrated that there is no single optimum or maximum permissible rate for subtitles. These rates are highly dependent on the type of content and what feels natural.

Broadcasters receive complaints about the rate of subtitles and part 1 of this study shows people have a consistent perception of what content is too slow and too fast. If perceived rate is not caused by actual rate, as shown in part 2, then what is it caused by? In part 1 of this study, it was likely caused by a sense of what is a natural rate. The fast/slow speech in these clips felt oddly fast or slow. The fast speech of a malfunctioning robot in a movie is intentionally odd. But the fast speech of a frightened character feels right. Secondly, people may identify hard to follow content as too fast/slow. In the case of subtitles with delay or errors, it becomes harder to match audio/lip-reading with the subtitles necessitating higher concentration. This may be exacerbated by high information density at high rates. Conversely, a sentence that crosses the boundary between two slow subtitles will also require effort to hold the first part of the sentence long enough to combine it with the second and make sense of it as whole. It is also possible previous studies have failed to make the meaning of their question clear and clarify the answers of participants. This study identified multiple cases of participants using the term "too slow" to describe increased latency. Complaints of live subtitles being "too fast" may be explained by a combination of latency, errors and necessitated editing all requiring high concentration as well as subtitling systems causing subtitles to "bunch up" and be played out at inconsistent rates. It should be noted that the use of these terms was clarified with participants in this study.

Previous studies and guidelines have insisted relatively low rates are needed to enable viewers to follow the content - even if they request otherwise. This study has shown that low rates are not required for viewers to feel that they are following the content sufficiently.

Conclusion

This study has shown that the perceived rate of subtitles for frequent users tends to align with those of speech for the hearing. It has shown that different content feels right at different speeds. Furthermore, the perceived rate of subtitles is not representative of the actual speed but is a symptom of technical issues and the overall natural feel of the programme. To avoid perceived issues with rate, subtitles should match the speech in timing and wording. We found no problems associated with the rate of subtitles when they matched natural speech, regardless of the rate in words per minute.

Acknowledgements

The author would like to thank Annabel Tiffin, Michelle Mayman, Jane Smith and all the BBC North West Tonight team for helping create realistic news clips that added so much to this study; to Colin Warhurst and Ahmed Razek at the BBC Blue Room for lending space for extra work to take this study a little further; to Gareth Ford Williams for providing resources to make this study happen; and to Mike Crabb for his guidance in test design and data analysis making this study the best it could be.

References

[1] OFCOM, 'The quality of live subtitling - Improving the viewer experience', 2013

[2] MIKUL C.: 'Caption quality: international approaches to standards and measurement', 2014

[3] JENSEMA C.: 'Viewer reaction to different television captioning speeds', *American Annals of the Deaf*, 1998, Vol. 143, pp. 318–324

[4] BURNHAM D., *ET AL*.: 'Parameters in television captioning for deaf and hard-of-hearing adults: Effects of caption rate versus text reduction on comprehension', *Journal of Deaf Studies and Deaf Education*, 2008, Vol. 13, pp. 391–404

[5] ROMERO-FRESCO P.: 'Standing on quicksand: Viewers' comprehension and reading patterns of respoken subtitles for the news', *Approaches to Translation Studies*, 2010, Vol. 32

[6] ROMERO-FRESCO P.: 'Quality of live subtitling: the reception of respoken subtitles in the UK', *Approaches to Translation Studies*, 2012, Vol. 36

[7] STEINFELD A.: 'The benefit to the deaf of real-time captions in a mainstream classroom environment', 1999

[8] JENSEMA C., MCCANN R., RAMSEY S.: 'Closed-caption television presentation speed and vocabulary', *American Annals of the Deaf*, 1996, Vol. 141, pp. 284–292

[9] NOLAND K.C., TRUONG L.H.: 'A survey of UK television viewing conditions', *BBC R&D - WHP 287*, 2015

[10] Action on Hearing Loss. Statistics. Action on Hearing Loss. http://www.actiononhearingloss.org.uk/your-hearing/about-deafness-and-hearing-loss/statistics.aspx Accessed: 28 April 2015

Introduction to *Electronics Letters*

Last year *Electronics Letters* celebrated its 50th year of publication. Launched in 1965, just two years before the first IBC, over the last five decades *Electronics Letters* has published over 43,000 papers and seen its scope evolve to reflect the amazing changes and advances in electronics since the 1960s.

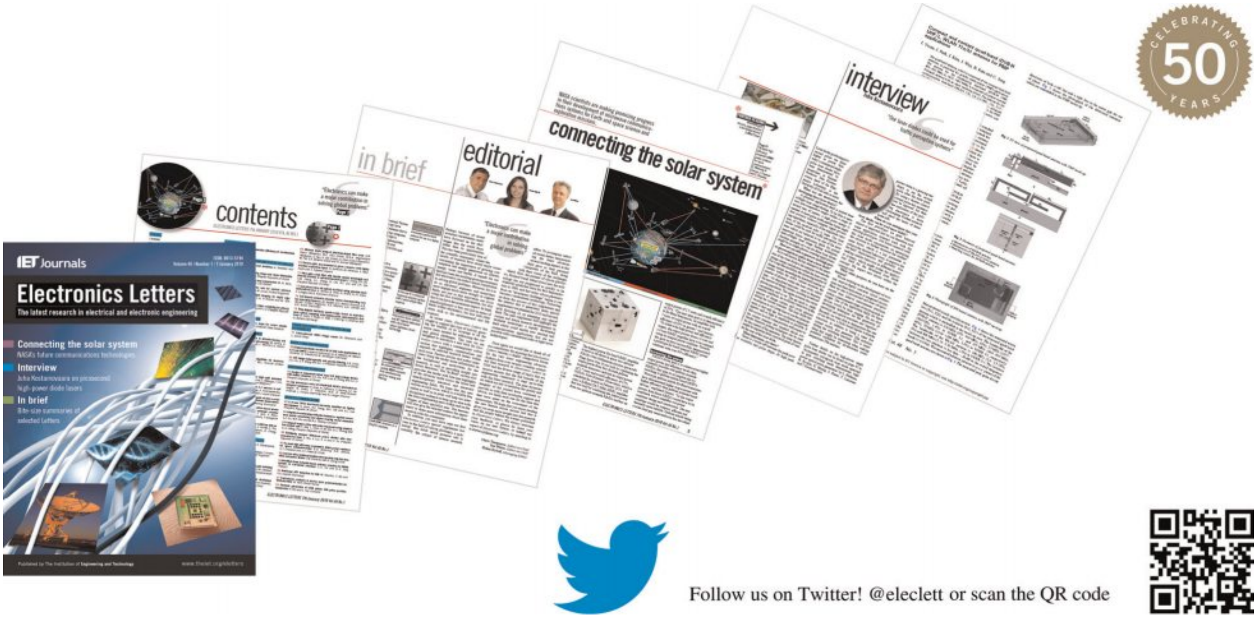
Electronics Letters[1] is a uniquely multidisciplinary rapid publication journal with a short paper format that allows researchers to quickly disseminate their work to a wide international audience. The broad scope of *Electronics Letters* encompasses virtually all aspects of electrical and electronic technology from the materials used to create circuits, through devices and systems, to the software used in a wide range of applications. The fields of research covered are relevant to many aspects of multimedia broadcasting including fundamental telecommunication technologies and video and image processing.

Each issue of *Electronics Letters* includes a magazine style news section. The news section includes feature articles based on some of the best papers in each issue, providing more background and insight into the work reported in the papers, direct from the researchers.

We hope you will enjoy reading the selection of papers and features included in this year's Best of the IET and IBC as examples of our content, and if you like what you read, all our feature articles are available for free via our web pages[1].

The *Electronics Letters* editorial team

[1] www.theiet.org/eletters


IET Journals
Electronics Letters
The latest research in electrical and electronic engineering
Connecting the solar system
Interview
In brief
contents
in brief
editorial
connecting the solar system
interview
CELEBRATING 50 YEARS

Follow us on Twitter! @eleclett or scan the QR code

Thumbnail extraction for HEVC using prediction modes

Wonjin Lee[1] *Gwanggil Jeon*[2] *Jechang Jeong*[1]

[1] *Department of Electronics and Computer Engineering, Hanyang University, Republic of Korea*
[2] *Department of Embedded Systems Engineering, Incheon National University, Republic of Korea*
E-mail: jjeong@hanyang.ac.kr

Abstract: The existing thumbnail extraction method generates a thumbnail by reducing the reconstructed frame after a full decoding process. This method is complex and requires considerable time, particularly for higher resolution video. To alleviate these issues, a fast thumbnail extraction method which utilises an intra-prediction mode for high-efficiency video coding (HEVC) is presented. The proposed method reconstructs only the 4 × 4 boundary pixels needed for a thumbnail image based on the intra-prediction mode. Experimental results indicate that the proposed method significantly reduces the computational complexity and extraction time for a thumbnail image. The visual quality of thumbnail images obtained with this method does not differ significantly from that of images extracted after a full decoding process.

Introduction

Developments in digital broadcasting technology and display devices have driven the need for ultra-high definition (UHD) video. New video compression techniques aid in devices efficiently saving or transmitting the significant amount of data that is essential for UHD video. To address this need, the high-efficiency video coding (HEVC) standard was designed by the video coding experts group and the moving picture experts group in January 2013 [1–5]. Various service applications using the HEVC standard have been developed and fast image processing methods for UHD are actively being investigated. One of these regarding the reduction of image size is called the thumbnail image. Thumbnail images are used for the fast indexing or searching of video, because they contain the overall elements which are needed to roughly represent the characteristics of the image. In addition, extracting a thumbnail image requires less memory and computational complexity than a complete reconstruction of the original image. In this Letter, we propose a fast thumbnail extraction method for HEVC, which significantly reduces the computational complexity required for thumbnail extraction.

Proposed algorithm

The main goal of our proposed method is to produce a thumbnail image without the use of a full decoding process (including in-loop filtering [4, 5]). The proposed method uses a sub-sampling method to extract thumbnails. The sub-sampling method extracts a representative pixel in each 4 × 4 area that is the summation of the residual sample and the prediction sample through the use of inverse transform and intra-prediction, respectively. The reconstructed pixel **rec** is defined as follows

$$\mathbf{rec}(x, y) = \mathbf{res}(x, y) + \mathbf{pred}(x, y) \tag{1}$$

where **rec** is the reconstructed pixel of the TU (transform unit) size, **res** and **pred** are residual samples by inverse transform and prediction samples by intra-prediction, respectively. The indices, x and y, are the sample position of each TU size, and the sub-sampling is performed in each 4 × 4 size. For example, when the TU size is 8 × 8, the sub-sampling is performed to $\{(x, y)\} = \{(3, 3), (7, 3), (3, 7), (7, 7)\}$.

The prediction block is produced using the upper and left reference pixels, which are the boundary pixels of the previously-reconstructed blocks according to intra-prediction modes [3]. The current intra-prediction

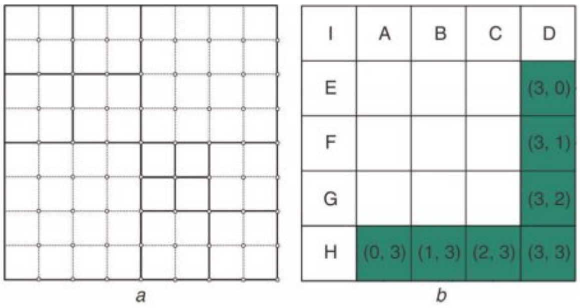


Figure 1 *Partial reconstruction for dash line*
a Sub-sampling position of 4 × 4 boundaries
b Partial reconstruction boundary of 4 × 4 TU

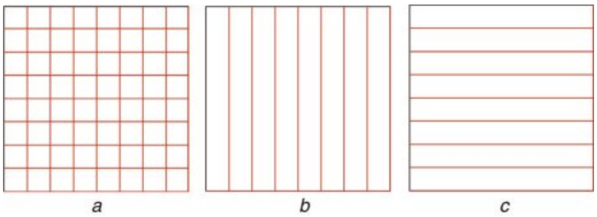


Figure 2 *Reconstruction lines of 32 × 32 TU according to intra-prediction mode*

a Reconstruction lines for intra-prediction modes (0, 1, 11–25)
b Reconstruction lines for intra-prediction modes (2–10)
c Reconstruction lines for intra-prediction modes (26–34)

generates all samples in a prediction block; however, the thumbnail image does not require all samples of the prediction block. Thus, our proposed method reconstructs only 4 × 4 boundaries that represent the samples needed for a thumbnail, as shown in Fig. 1. The solid lines, dashed lines and small ellipses in Fig. 1*a* are TU boundaries, 4 × 4 boundaries and sub-sampling positions, respectively. Partial intra-prediction of the proposed method produces only dashed lines (4 × 4 boundaries). **res** in can be obtained by performing IDCT as follows

$$\mathbf{res} = \boldsymbol{C}^{\mathrm{T}} \cdot \mathbf{RES} \cdot \boldsymbol{C} \tag{2}$$

where **RES** is the DCT-transformed **res**, ***C*** is the integer DCT basis with sizes ranging from 4 × 4 to 32 × 32 and $\boldsymbol{C}^{\mathrm{T}}$ is the transposed ***C***. The proposed method can reduce the computational complexity of by obtaining only 4 × 4 boundary samples. When TU is 4 × 4, the boundary positions are $\{(x, y)\} = \{(0, 3), (1, 3), (2, 3), (3, 0), (3, 1), (3, 2), (3, 3)\}$. As shown in Fig. 1*b*, a partial IDCT for the 4 × 4 boundary samples is obtained as follows

$$\mathbf{res}_{\mathrm{ver}} = \boldsymbol{C}^{\mathrm{T}} \cdot \mathbf{RES} \cdot \boldsymbol{C}_{\mathrm{ver}}, \quad \mathbf{res}_{\mathrm{hor}} = \boldsymbol{C}_{\mathrm{ver}}^{\mathrm{T}} \cdot \mathbf{RES} \cdot \boldsymbol{C} \tag{3}$$

where $\mathbf{res}_{\mathrm{ver}}$ is the 4 × 1 residual samples of the vertical boundary, $\mathbf{res}_{\mathrm{hor}}$ is the 1 × 4 residual samples of the horizontal boundary, and $\boldsymbol{C}_{\mathrm{ver}}$ is the 4 × 1 matrix of the 4th row of ***C***. For instance, when TU and PU (prediction unit) are 32 × 32, the partial intra-prediction produces only 4 × 4 boundary samples (448 samples) and the eight $\boldsymbol{C}_{\mathrm{ver}}$, which are 32 × 1 matrices of the 4th, 8th, 12th, 16th, 20th, 24th, 28th, and 32th rows of ***C***, and are

Table 1 Consumed time comparison (in seconds)

Class	Sequence	Method [6]	Prop.	Time reduction (%)
Class *A*	PeopleOnTheStreet (POS)	49	33	33
	Traffic (TR)	46	32	30
Class *B*	BQTerrace (BQT)	95	67	29
	Cactus (CA)	75	52	31
	Kimono (KI)	29	20	31
	ParkScene (PS)	38	27	29
Class *C*	BasketballDrill (BD)	17	12	29
	BQMall (BQM)	20	15	25
	PartyScene (PS)	24	18	25
Class *D*	BasketballPass (BP)	5	4	20
	BlowingBubbles (BB)	7	5	29
	BQSquare (BQS)	7	6	14
	RaceHorses (RH)	3	2	33
Class *E*	FourPeople (FP)	39	27	31
	Johnny (JO)	30	20	33
	KristenAndSara (KAS)	33	22	33
Class *F*	BasketballDrillText (BDT)	18	12	33
	ChinaSpeed (CS)	32	22	31
	SlideEditing (SE)	26	19	27
	SlideShow (SS)	22	14	36

Table 2 Computational complexity comparison for one thumbnail pixel

Class	Sequence	Method [6]			Prop.		
		Add.	Mult.	Comp.	Add.	Mult.	Comp.
Class *A*	POS	457	213	41	207	119	15
	TR	511	256	34	261	161	12
Class *B*	BQT	221	106	19	101	60	7
	CA	244	123	17	123	76	6
	KI	277	160	10	168	113	3
	PS	247	122	17	128	77	6
Class *C*	BD	45	20	4	19	11	1
	BQM	45	20	4	20	12	1
	PS	42	16	6	17	8	2
Class *D*	BP	12	5	0	6	3	0
	BB	10	4	1	4	2	0
	BQS	9	4	1	4	2	0
	RH	11	5	1	5	3	0
Class *E*	FP	103	50	7	40	29	2
	JO	91	48	5	45	29	1
	KAS	102	56	6	52	34	2
Class *F*	BDT	44	19	4	18	10	1
	CS	88	44	7	42	26	3
	SE	72	30	12	22	12	4
	SS	57	26	5	21	11	1

obtained via partial IDCT. To further reduce the computational complexity, the proposed method uses the intra-prediction mode. There are 35 intra-prediction modes for the HEVC (mode 0 to mode 34) [3]. Mode 0 is a planar mode, mode 1 is a DC mode, and modes 2 to 34 are angular modes. The 4 × 4 boundary becomes the reference pixels of the next coding blocks, and the number of reference pixels for the intra-prediction differs in each mode. When the intra mode ranges (0, 1, 11–25), the reference pixels need vertical and horizontal lines, when the intra mode ranges from 2 to 10, reference pixels only need vertical lines, and when the intra mode ranges from 26 to 34, the reference pixels only need horizontal lines, as shown in Fig. 2. Therefore, the computational complexity can be further reduced by reconstructing only vertical or horizontal lines based on the intra-prediction mode.

Experimental results

The proposed method was compared with the conventional algorithm, which is a 4 × 4 sub-sampling of the reconstructed image from the HEVC reference decoder [6]. The proposed algorithm was tested on HEVC video reference software (HM 13.0). All frames are type I frames, and all test sequences of the HEVC were tested. The test sequences employed were classified into the following six classes: Class *A* (2560 × 1600), Class *B* (1920 × 1080), Class *C* (832 × 480), Class *D* (416 × 240), Class *E* (1280 × 720) and Class *F* (832 × 480, 1024 × 768, 1280 × 720).

Table 1 shows the time comparison. The proposed algorithm can reduce the thumbnail extraction time by 14 to 36% with respect to conventional methods. Table 2 shows the computational complexity of one pixel of the thumbnail image compared with the conventional method, as well as the operations of addition, comparison, and multiplication that are required per thumbnail pixel. The computational complexity was significantly reduced in all sequences of the proposed method, as compared with conventional methods. In terms of the computational complexity of the UHD POS sequence of Class *A*, the addition, multiplication, and comparison operations were reduced by 55%, 44%, and 63%, respectively. A visual quality comparison of the PeopleOnTheStreet sequence can

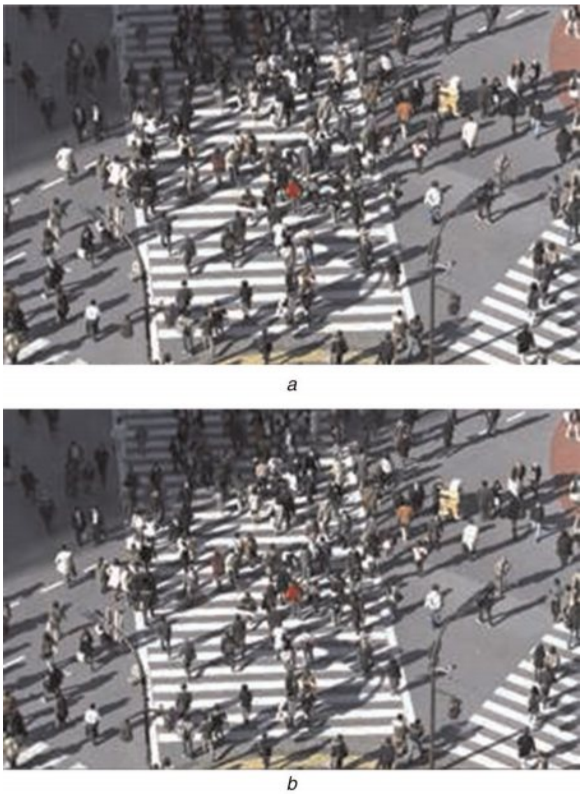


Figure 3 *Subjective result of thumbnail image (640 × 400) of PeopleOnTheStreet sequence (2560 × 1600)*

a Method [6]
b Proposed method

be seen in Fig. 3. The subjective qualities of both results are very similar despite the considerable reduction in thumbnail extraction time and computational complexity.

Conclusions

In this Letter, a fast thumbnail extraction method utilising the intra-prediction mode is proposed to reconstruct only the 4 × 4 boundary. The proposed method can reduce computational complexity and thumbnail extraction times while maintaining visual quality. Experimental results show that the proposed algorithm achieved extraction time reductions ranging from 14% to 36%, in addition to a significant reduction in computational complexity. In terms of the subjective result, the proposed method generated images that do not significantly vary from those generated via the full decoding of a conventional algorithm.

Acknowledgments

This work was supported by the Brain Korea 21 plus (BK 21 plus) project in 2014 and by the NRF of Korea Grant funded by the Korean Government (MSIP)(2014025627).

References

[1] SULLIVAN G.J., OHM J.R., HAN W.J., WIEGAND T.: ‘Overview of the high efficiency video coding (HEVC) standard’, *IEEE Trans. Circuits Syst. Video Technol.*, 2012, Vol. 22, No. 12, pp. 1649–1668

[2] KIM I.K., MIN J., LEE T., HAN W.J., PARK J.: ‘Block partitioning structure in the HEVC standard’, *IEEE Trans. Circuits Syst. Video Technol.*, 2012, Vol. 22, No. 12, pp. 1697–1706

[3] LAINEMA J., BOSSEN F., HAN W.J., MIN J., UGUR K.: ‘Intra coding of the HEVC standard’, *IEEE Trans. Circuits Syst. Video Technol.*, 2012, Vol. 22, No. 12, pp. 1792–1801

[4] NORKIN A., BJØNTEGAARD G., FULDSETH A., NARROSCHKE M., IKEDA M., ANDERSSON K., ZHOU M., AUWERA , VAN DER AUWERA G.: ‘HEVC deblocking filter’, *IEEE Trans. Circuits Syst. Video Technol.*, 2012, Vol. 22, No. 12, pp. 1746–1754

[5] FU C.M., ALSHINA E., ALSHIN A., HUANG Y.W., CHEN C.Y., TSAI C.Y., HSU C.W., LEI S.M., PARK J.H., HAN W.J.: ‘Sample adaptive offset in the HEVC standard’, *IEEE Trans. Circuits Syst. Video Technol.*, 2012, Vol. 22, No. 12, pp. 1755–1764

[6] Flynn, D.: H.265/HEVC Reference Software (online), https://hevc.hhi.fraunhofer.de/trac/hevc/browser/branches/HM-13.0-dev, March 2013

Selected content from Electronic Letters

Multiuser motion recognition system using smartphone LED luminescence

Byung-Hun Oh[1] *Jung-Hyun Kim*[3] *Kwang-Woo Chung*[2] *Kwang-Sook Hong*[1]

[1]*School of Information and Communication Engineering, Sungkyunkwan University, 300, Chunchun-dong, Jangangu, Suwon, KyungKi-do 440-746, Republic of Korea*
[2]*Department of Railway Operation System Engineering, Korea National University of Transportation, 157, CheoldobangmulgKwan-ro, Uiwang-si, Kyungki-do 437-763, Republic of Korea*
[3]*Central Technology Appraisal Institute, KOTEC, 11th Songdo-Centrod, Songdo-dong, Yeonsu-gu, Incheon 406-840, Republic of Korea*
E-mail: sincelife@skku.edu

Abstract: A novel multiuser motion recognition system that uses a smartphone detection method and tracking interface is proposed. This multiuser motion recognition system provides a motion control interface between devices using an image-based object tracking algorithm through smartphone light-emitting diode (LED) luminescence. It requires just a universal serial bus (USB) camera and personal smartphone instead of expensive equipment to provide a motion control interface. It recognises the user's gestures by tracking three-dimensional (3D) distance and the rotation angle of the smartphone, which acts essentially as a controller in the user's hand.

Introduction

As the use of smartphones becomes even more widespread, with features such as TV–mobile device interaction, the case for using the device in convergence with other devices is increasing. Accordingly, an easy interworking function between devices provides improved user convenience, although a more readily accessible device that enables high-accuracy interaction technologies needs to be developed [1].

Until the mid-2000s, users had to use tools such as controllers and keyboards to interact with devices. These tools had several buttons that the user could press, so they sent the user's commands to the device by electronic signals or radio waves. Then, since 2006, a new type of controller began to be released. These controllers had fewer buttons on them, but more importantly, they had special sensors, such as accelerometers, gyroscopes and infrared sensors. Since 2010, interface technologies that do not require controllers have been developed. Instead, they used methods of detecting human body positions from incoming images, using special cameras. The motion control industry is growing at an extremely fast pace, and technologies used in motion control are becoming more important than ever [2, 3]. However, these interface technologies have some disadvantages. They use many hardware sensors to provide accurate user interfaces, but this leads to increases in the price of the product. Moreover, to detect the human body in a three-dimensional (3D) environment, they use special RGB cameras and depth sensors that are still quite expensive. These additional equipment items increase the cost, and become a financial burden to some people who want to enjoy motion control [4].

In this Letter, a novel multiuser motion recognition system that uses smartphone light-emitting diode luminescence is proposed. This multiuser motion recognition system provides a motion control interface between devices using only a universal serial bus (USB) camera and a smartphone. It can recognise a user's gestures by tracking the 3D position and the angle of the smartphone through the camera image while the user holds a smartphone in his/her user hand.

System concept

Fig. 1 shows the concept for a novel multiuser motion recognition system.

In the first step, the system accurately detects multiple light-emitting smartphones in the image taken from the camera of the smart device (e.g. desktop PC or smart TV) for interaction. Then, the camera takes images of the user and starts to search for the location of the smartphone(s) using image-processing algorithms. The motion recognition for the interaction is recognised in terms of smartphone coordinates and area in the second step. Finally, with the estimated data, additional commands can

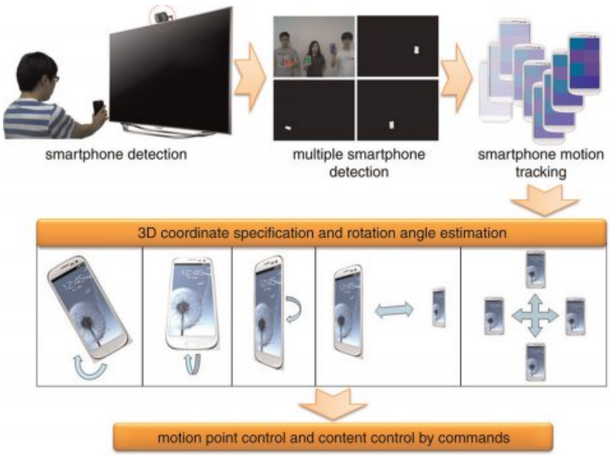

smartphone detection
multiple smartphone detection
smartphone motion tracking
3D coordinate specification and rotation angle estimation
motion point control and content control by commands

Figure 1 *Concept of suggested system*

be prepared. Moreover, using the smartphone as a controller between the smart devices, an excellent interface can be created.

Real-time multiple smartphone detection

The multiple smartphone detection involves six major steps on the basis of two different colour models, the YCbCgCr colour model and HSV together with the colour model conversion process performed in parallel: (i) in a normalisation process step, the colour values are divided by the sum of the three components of the RGB to make it robust to illumination change; (ii) the conversion step of the YCbCgCr colour model separates the colour difference information (Gb, Cg, Cr) except for the brightness component in the normalised RGB colour model, and extracts the adaptive threshold values; (iii) a noise removal step, after performing a binarisation operation, removes the noise using erosion and expansion operations; (iv) in the conversion step of the HSV colour model, the normalised RGB colour model is converted to an HSV colour model, and separated into each of the channels (hue, saturation, value); (v) in the mask image generation step, a prominent colour is extracted from the saturation and value channel using the adaptive threshold, and by performing an 'AND' operation with the saturation channel and the value channel generated and (vi) finally, the system tracks the target object in the result image, created by computing the AND operation with the mask image of the HSV colour model and the result image of the YCbCgCr colour model.

Smartphone tracking interface

In this Letter, the smartphone tracking interface can be divided into three aspects:

(i) 2D pixel coordinate estimation is estimated using the vertex coordinates of the detected rectangle. When the detection algorithm detects the object, four vertex coordinates can be acquired. Equations for estimating the x and y pixel coordinates are listed below

$$
\begin{aligned}
X\_\mathrm{PIXEL} &= \frac{\mathrm{Point\ 1.}x + \mathrm{Point\ 2.}x + \mathrm{Point\ 3.}x + \mathrm{Point\ 4.}x}{4}\\
Y\_\mathrm{PIXEL} &= \frac{\mathrm{Point\ 1.}y + \mathrm{Point\ 2.}y + \mathrm{Point\ 3.}y + \mathrm{Point\ 4.}y}{4}
\end{aligned}
\tag{1}
$$

X_PIXEL and Y_PIXEL indicate the x and y coordinates of the smartphone detected in pixels. Point1.x, 2.x, 3.x and 4.x and Point1.y, 2.y, 3.y and 4.y indicate the x and y coordinates of each vertex point.

(ii) 3D distance estimation is used regarding the width of the smartphone detecting rectangle. The equation and variables used for estimating the z-axis distance are as follows:

$$
Z\mathrm{DISTANCE}= \frac{\mathrm{DEIVCEWIDTH}\times\mathrm{CAMERAWIDTH}}{\mathrm{WidthLength}\times\tan(\mathrm{WIDTHFOV}/2)\times 2}
\tag{2}
$$

Z_DISTANCE indicates the distance between the camera and smartphone (cm). DEVICE_WIDTH is the real width of the smartphone. CAMERA_WIDTH is the width of the image taken by the camera (pixels). Width_Length indicates the width of the smartphone detection rectangle (pixels) and WIDTH_FOV is the field of view of the camera.

(iii) Rotated angle estimation is estimated using the variation of width and height of the smartphones detected. The width of the detecting rectangle was used to estimate the y-axis rotation angle and the height of the detecting rectangle was used for estimating the x-axis rotation angle. Equations and variables for estimating the x and y axes rotation angles are as follows:

$$
\begin{aligned}
X\_\mathrm{ANGLE} &= \frac{\mathrm{Height\_Length}}{\mathrm{Max\_Height\_Length}} \times X\_\mathrm{MaxAngle}\\
Y\_\mathrm{ANGLE} &= \frac{\mathrm{Width\_Length}}{\mathrm{Max\_Width\_Length}} \times Y\_\mathrm{MaxAngle}
\end{aligned}
\tag{3}
$$

X_ANGLE indicates the x-axis rotation angle (θ), and Height_Length is the height of the smartphone detected (pixels). Max_Height_Length is the maximum height of the smartphone detected, which can be acquired when the smartphone screen faces straight at the camera. X_MaxAngle is the maximum angle that the device can rotate in the x-axis. Y_ANGLE is the same except change the X_XANGLE equation from height to width. To estimate the z-axis rotation angle, we used the three points of the detected rectangle: two vertex points and the central

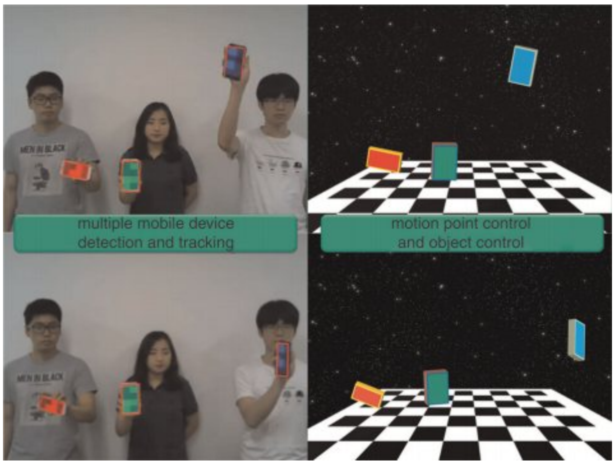

multiple mobile device detection and tracking
motion point control and object control

Figure 2 *Example of manipulating object by applying mobile device tracking interface*

point. An example of manipulating the object by applying the smartphone tracking interface is shown in Fig. 2.

Experiments and performance evaluation

We prepared five kinds of videos by adjusting the distance between the camera and the user's feet from 100 to 300 cm at intervals of 50 cm to check the performance of the detection algorithms. We also prepared three different databases at each distance by changing the displayed colours on the smartphone screen to red, blue and green. We tagged smartphone location coordinates at each frame of the database, and saved smartphone location coordinates to files by clicking the white dot at the centre of the device screen.

We measured the smartphone detection rate using the prepared video databases. We measured the size of the smartphone at five distances. We set the range of successful detection using the size of the smartphone, and decided whether the detection was successful by checking whether the detected point was in the range. Table 1 shows the detection rates of the smartphone by colour and distance.

Table 1 Experimental results of smartphone detector

	100 cm	150 cm	200 cm	250 cm	300 cm
Red	93.1%	91.7%	89.5%	84.4%	82.4%
	(931/ 1000)	(917/ 1000)	(895/ 1000)	(844/ 1000)	(824/ 1000)
Green	88.7%	90.1%	91.2%	92.7%	92.1%
	(887/ 1000)	(901/ 1000)	(912/ 1000)	(927/ 1000)	(921/ 1000)
Blue	93.4%	91.2%	92.5%	91.4%	92.5%
	(934/ 1000)	(912/ 1000)	(925/ 1000)	(914/ 1000)	(925/ 1000)

Table 2 Experimental results of estimated pixels and distance

	x pixel	*y* pixel	*z* distance
Red	2.51% (1000 frames)	1.57% (1000 frames)	3.12% (1000 frames)
Green	1.97% (1000 frames)	1.04% (1000 frames)	2.77% (1000 frames)
Blue	1.1% (1000 frames)	0.71% (1000 frames)	2.43% (1000 frames)

Table 3 Experimental results of estimated rotation angle

	x angle	*y* angle	*z* angle
Red	1.24% (1000 frames)	1.64% (1000 frames)	0.76% (1000 frames)
Green	0.71% (1000 frames)	1.41% (1000 frames)	0.77% (1000 frames)
Blue	0.79% (1000 frames)	0.82% (1000 frames)	0.62% (1000 frames)

With blue light emission, the detection rate was 92.20% on average, the highest among the three colours. Overall, the average was 90.46%, and the performance time was 43.2 ms per frame.

We also measured the error rates for the estimated *x* and *y* pixels and *z* distances. The same databases were used, and we used tagged coordinates and estimated coordinates in calculating error rates. In the experimental results, the error rates (*x* pixel, *y* pixel and *z* distance) were 1.87, 1.92 and 2.77%, respectively. Overall, the average was 2.18%, and Table 2 shows the results of the experiment.

Finally, we measured the error rates for the estimated rotation angle. The experiment for the rotation angle estimation used 1000 frames for angles ranging from 08 to 1608 at intervals of about 108. In the experimental results, the error rates (*x* angle, *y* angle and *z* angle) were 0.91, 1.29 and 0.71%, respectively. Overall, the average was 0.97%. Table 3 shows the results of the experiment.

Conclusion

In contrast to other motion-based recognition methods, the proposed method is unique in three respects. First, the proposed multiple smartphone detection scheme shows good robustness against lighting variance and complex backgrounds using the HSV colour model and the

YCbCgCr colour model. Secondly, the proposed system can support user interfaces between devices using only a USB camera and a smart device. Finally, the proposed method enables a smartphone to serve as a motion controller, so that a user can enjoy a tangible interaction without the need to buy another device, so it is possible to provide a more convenient and efficient user environment. Since the user only needs a computer, webcam and smartphone, this multiuser motion recognition system is economical compared with other motion recognition systems. In our experiments, the method showed a high recognition rate and the system can serve as a prototype for the motion recognition system.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2014023704) and was also supported by the Basic Science Research Program through NRF of Korea, funded by the MOE (NRF-2010-0020210).

References

[1] BALDAUF M., FROHLICH P.: 'The augmented video wall: multi-user AR interaction with public display'. Proc. CHI, Paris, France, May 2013, pp. 3015–3018

[2] SUNG J., PONCE C., SELMAN B., SAXENA A.: 'Human activity detection from RGBD images'. AAAI 2011, Workshop, San Francisco, CA, USA, August 2011, pp. 47–55

[3] LI W., ZHANG Z., LIU Z.: 'Action recognition based on a bag of 3d points'. CVPRW, San Francisco, CA, USA, June 2010, pp. 9–14

[4] OIKONOMIDIS I., KYRIAZIS N., ARGYROS A.A.: 'Efficient model-based 3D tracking of hand articulations using kinect'. British Machine Vision Conf., Dundee, Scotland, September 2011, pp. 101.1–101.11

Selected content from Electronic Letters

Synthesising frontal face image using elastic net penalty and neighbourhood consistency prior

Yuanhong Hao[*] *Chun Qi*

School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China
[*]Also with the North Automatic Control Technology Institute, Taiyuan, People's Republic of China
E-mail: facerecognition@163.com

Abstract: Traditional frontal face image synthesis based on the ℓ_1-penalty has achieved remarkable success. However, the ℓ_1-penalty on reconstruction coefficients has the drawback of instability when processing high-dimensional data (e.g. a facial image including hundreds of pixels). Moreover, the traditional ℓ_1-penalty-based method requires consistency between the corresponding patches in frontal and profile faces, which is hard to guarantee due to self-occlusion. To overcome the instability problem of the traditional method, an extension of the ℓ_1-penalty-based frontal face synthesis method, which benefits from the nature of the elastic net, is presented. 3 addition, to enhance the aforementioned consistency, a neighbourhood consistency penalty is imposed onto the reconstruction coefficients using the connected neighbour patches of the current patch. Furthermore, to ensure the synthesised result faithfully approximates the ground truth, a sparse neighbour selection strategy is introduced for finding related neighbours adaptively. Experimental results demonstrate the superiority of the proposed method over some state-of-the-art methods in both visual and quantitative comparisons.

Introduction

Given only a photograph of a person's profile face, can we infer how the face might look from a frontal viewpoint? This type of problem often occurs in facial animation systems, video surveillance and face recognition across a pose. Frontal face synthesis refers to synthesising one person's virtual frontal images given his or her face images at other poses. For human faces, which share a highly similar structure, the human brain learns the relationship between the profile and frontal images through experiences of facial images at different orientations [1]. Hence, a reasonable solution is to find the relationship through statistic learning-based approaches.

Taking advantage of the statistic learning-based point of view, Chai *et al.* [2] proposed an ordinary least squares (OLS)-based method. However, OLS does not generalise well when the linear system is ill-conditioned. To solve the ill-posed problem of face synthesis, Zhang *et al.* [3] and Zhao *et al.* [1] successfully synthesised face images via ℓ_1-penalised least squares regression, which is also known as the least absolute shrinkage and selection operator (LASSO) [4]. However, LASSO is unstable when the predictors are highly correlated. Especially, for the high-dimensional data, which frequently exists in face image processing applications, sample correlation can be large even when predictors are independent [4]. To overcome this problem, the elastic net (ENet) [4], which is a compromise between the LASSO penalty and ridge penalty, is an effective method. Therefore, it is conceivable that, if we impose an ENet penalty into the coefficients, the advantage of the statistic learning-based approach can be reasonably highlighted. Additionally, these traditional methods require consistency between the corresponding patches in frontal and profile faces, which is hard to guarantee by just using a coarse alignment [1, 5]. To tackle this problem, Zhao *et al.* [1] employed a triangulation-based partition criterion. Different from their strategy, we enhance the above consistency by regularising the solution using the connected neighbour patches of the current patch.

ENet penalty-based frontal face synthesis

Based on the motivation stated above, we propose the following method. Suppose we have the training datasets $\boldsymbol{D}_0 \in \Re^{d\times n}$ and $\boldsymbol{D}_p \in \Re^{d\times n}$, where the columns of $\boldsymbol{D}_0$ and $\boldsymbol{D}_p$ are, respectively, composed of the vectorised frontal facial images and the corresponding profile faces under pose p. Given an input image $\boldsymbol{y}_p$ whose pose is p, we divide $\boldsymbol{y}_p$ into s small uniform patches in raster-scan order to form the test set $\{y_{(j,p)}\} \in \Re^{b \times s}$ with $j = 1, \ldots, s$. The same operation is carried out on the images in D_p and D_0. Then, synthesising the corresponding jth virtual frontal patch $y_{(j,0)}$ for the jth profile patch $y_{(j,p)}$ follows two steps. First, we obtain the reconstruction coefficients by

$$\widehat{v}_j = \arg\min_{v_j} \left\| y_{(j,p)} - D_{(j,p)} v_j \right\|_2^2 + l_j \left\{ \left\| v_j \right\|_2^2 + a \left\| v_j \right\|_1 \right\} \tag{1}$$

where $v_j = (v_j^1, \ldots, v_j^i)^{\mathrm{T}} \in \Re^{n \times 1}$ is the reconstruction coefficient vector; l_j $(l_j \geq 0)$ denotes the regularisation parameter and the constant a $(a \geq 0)$ controls the compromise between LASSO and ridge regression. Secondly, the frontal patch $y_{(j,0)}$ will be predicted via

$$y_{(j,0)} = D_{(j,0)} v_j \tag{2}$$

Neighbourhood consistency penalty

According to Fig. 1, small local patches at the same position of different poses, e.g. $y_{(j,p)}$ and $y_{(j,0)}$, are totally different and share few common facial textures. This is due to self-occlusion and pose variation. The inconsistency between $y_{(j,p)}$ and $y_{(j,0)}$ will lead to the inaccuracy of predicting $y_{(j,0)}$ from $y_{(j,p)}$. In comparison with $y_{(j,p)}$ and $y_{(j,0)}$, a larger patch $y_{(j,p)}$ at the jth position and its frontal counterpart $y_{(j,0)}$ share relatively more common facial textures (see Fig. 1). In other words, the consistency between $y_{(j,0)}$ and $y_{(j,p)}$ is more greater than the one between $y_{(j,p)}$ and $y_{(j,0)}$. Hence, if we regularise the coefficients vector v_j with some appropriate perturbation using $y_{(j,p)}$, then the inconsistency between $y_{(j,p)}$ and $y_{(j,0)}$ can be mitigated. To achieve this aim, we extract $y_{(j,p)}$ by the following strategy: for the jth profile vectored patch $y_{(j,p)}$ of the input image y_p, we collect its k connected neighbour patches in raster-scan order. Then, these connected neighbour patches and the centre patch $y_{(j,p)}$ together form $Y_{(j,p)} \in \Re^{m \times 1}$ in their original order. For the jth position of all the training images in D_p, we extract the corresponding larger patches in the same way until the dataset $D^*_{(j,p)} = \left\{ Y^i_{(j,p)} \right\}_{i=1}^n \in \Re^{m \times n}$ is constructed. Based on the above motivation, we add a new regularisation item into (1) as follows:

$$\widehat{v}_j = \arg\min_{v_j} \left\{ \begin{array}{l} \left\| y_{(j,p)} - D_{(j,p)} v_j \right\|_2^2 + l_j \left(\left\| v_j \right\|_2^2 + a \left\| v_j \right\|_1 \right) \\ + h^2 \left\| Y_{(j,p)} - D^*_{(j,p)} v_j \right\|_2^2 \end{array} \right\} \tag{3}$$

where h $(h \geq 0)$ is a regularisation parameter. The last term is named as the neighbourhood consistency penalty (NCR) in our method.

Equation (3) is equal to

$$\widehat{v}_j = \arg\min_{v_j} \left\{ \left\| \begin{bmatrix} y_{(j,p)} \\ hY_{(j,p)} \end{bmatrix} - \begin{bmatrix} D_{(j,p)} \\ hD^*_{(j,p)} \end{bmatrix} * v_j \right\|_2^2 + l_j \left(\left\| v_j \right\|_2^2 + a \left\| v_j \right\|_1 \right) \right\} \tag{4}$$

Equation (4) can be efficiently solved by the damped iterative thresholding algorithm [6].

Sparse neighbour selection

To enhance the similarity between a synthesised virtual face and the ground truth, we incorporate the sparse neighbour selection strategy [1] into our model in this Section. Once we obtain the coefficient vector v_j by optimising (4), we ensure the similarity constraint by the following updating criteria:

$$v_j^i = \begin{cases} v_j^i, & |v_j^i| \geq s, \\ 0, & \text{otherwise}, \end{cases} \quad 1 \leq i \leq n \tag{5}$$

According to [1], we set $s = 0.001$. In this way, the most similar patches can be selected for predicting the virtual frontal patch. The updated reconstruction coefficient vector is termed as v_j^*. Finally, we feed v_j^* into (6) to calculate the corresponding virtual frontal patch

$$y_{(j,0)} = D_{(j,0)} v_j^* \tag{6}$$

Choice of parameters

For choosing the regularisation parameter l_j adaptively, a five-fold cross-validation (CV) [4] is employed. We pick a relatively small grid of values for parameters $\{l_j^i\}_{i=1}^v$ for the jth patch, i.e. using the value in (0.6, 0.605, 0.61, 0.615, 0.62). We choose the optimal l_j^i that minimises the smallest CV error for the jth patch.

Summary of proposed method

Given an input image $y_p = \left\{ y_{(j,p)} \right\} \in \Re^{b \times s}$, we can feed its s small uniform patches into the algorithm, respectively, until all the corresponding frontal patches are generated. Then,

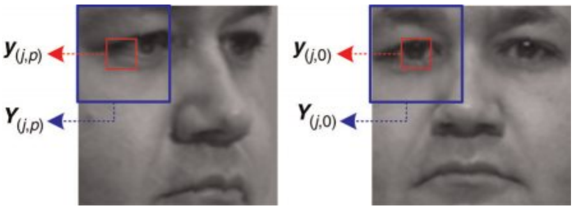


Figure 1 *Example showing lack of consistency between small local patch $y_{(j,p)}$ and corresponding frontal patch $y_{(j,0)}$*

all the virtual frontal patches are added into $\boldsymbol{T} = \left\{ \boldsymbol{y}_{(j,0)} \right\}_{j=1}^{s}$. The final virtual frontal face $\boldsymbol{y}_0$ will be produced by merging all the patches in $\boldsymbol{T}$. For the overlapping regions, the strategy in [7] is applied.

Experiments

Experiments are conducted on the MultiPIE database [8]. Seven pose subsets of the database are utilised, including the poses (frontal), 050 (+158), 041 (+308), 190 (+458), 140 (−158), 130 (−308) and 080 (−458). We take neutral expression and frontal lighting images from the session four, which has a total of 239 subjects. The first 100 subjects are taken as testing sets and the remaining 139 subjects are taken as training sets. All the faces are aligned by fixing the eye positions for the corresponding poses and cropped to 64 × 64 pixels.

After performing our experiments with a different patch size b and sampling step size, we find that the two sizes should be neither too large nor too small. Assigning a large size for the local patch may be inapposite because different persons have different local geometric shapes [2], while a too small patch size may result in lack of consistency between local patches of frontal and profile faces. Additionally, due to the nonlinearity and variety of high-dimensional input samples, the choice of neighbourhood size k usually affects the trade-off between the representation of the facial textures in $\boldsymbol{Y}_{(j,p)}$ and computational cost. Hence, by considering both the performance and time consumption, we set the sampling step size as 4 (refer to the sampling strategy in [2]) and k as 24. For the proposed approach (ENet-based face synthesis with NCR, EN with NCR), we set $b = 12 \times 12$, 16×16 and 24×24 for the viewpoint of + 158, + 308 and + 458, respectively, because the performance peaks when using these settings for the corresponding poses. As for parameters h and a, we set $h = 0.053$ and $a = 0.04$ empirically.

To demonstrate the effectiveness of EN with NCR, we compare our results with LLR [2], the local sparse representation-based method (LSRR) [3] and simple multiview face hallucination (SIMFH) [9]. Quantitative comparisons are provided based on the average peak signal-to-noise ratio (PSNR) values. The quantitative comparisons are shown in Fig. 2, where the values of PSNR (unit: dB) are averaged over the 100 inputs. According to the metric, it seems that we obtain a better performance than others. In addition, to verify the key role of NCR, we carry out the experiment using only (1) (i.e. EN without NCR) with three patch sizes in Fig. 2. The performance of EN with NCR is better than that without NCR. We think this is due to the loss of consistency of small patches, while large patches make the synthesising a nonlinear problem. In this case, linear regression is not appropriate for high-dimensional data. Thus, a combination of both is necessary. Visual comparisons are illustrated in Fig. 3, from which we can observe that our method produces the synthesised faces with the finest details. Obviously, Fig. 3 also demonstrates the superiority of our approach.

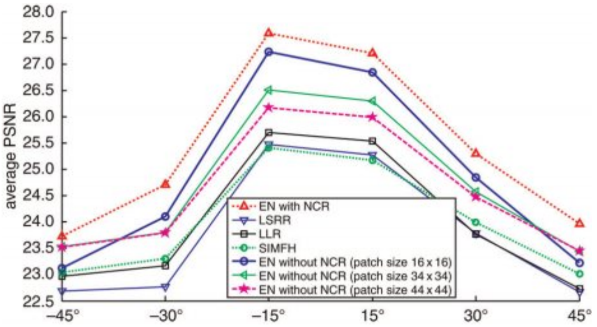


Figure 2 *Average PSNR values of results using different methods*

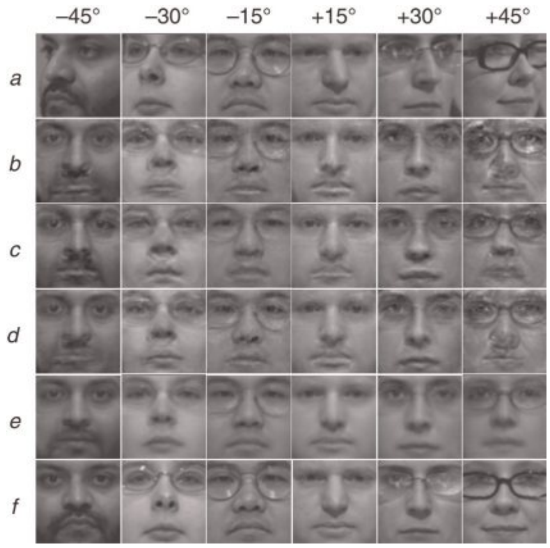


Figure 3 *Visual comparison results on MultiPIE database*

a Row 1 shows input profile faces with different poses
b Row 2 shows results using LLR
c Row 3 shows results for SIMFH
d Row 4 shows results using LSRR
e Row 5 gives our results
f Last row shows ground truth frontal faces

Conclusion

This Letter has presented a method for synthesising a virtual frontal face based on the ENet penalty. Additionally, we propose a new regularisation term to enhance the consistency between local patch pairs of frontal and profile faces. The proposed method can predict facial details effectively, which is demonstrated by the results of both visual and quantitative comparisons.

Acknowledgment

This work was supported by the Research Fund for the Doctoral Program of Higher Education of China (no. 20110201110012).

References

[1] ZHAO L., GAO X.B., YUAN Y., TAO D.P.: 'Sparse frontal face image synthesis from an arbitrary profile image', *Neurocomputing*, 2014, Vol. 128, pp. 466–475

[2] CHAI X., SHAN S., CHEN X., GAO W.: 'Locally linear regression for pose-invariant face recognition', *IEEE Trans. Image Process.*, 2007, Vol. 16, No. 7, pp. 1716–1725

[3] ZHANG H.C., ZHANG Y.N., HUANG T.S.: 'Pose-robust face recognition via sparse representation', *Pattern Recognit.*, 2013, Vol. 46, No. 5, pp. 1511–1521

[4] ZOU H., HASTIE T.: 'Regularization and variable selection via the elastic net', *J. R. Stat. Soc. Ser. B*, 2005, Vol. 67, No. 2, pp. 301–320

[5] SHARMA A., HAJ M.A., *ET AL.*: 'Robust pose invariant face recognition using coupled latent space discriminant analyses', *Comput. Vis. Image Underst.*, 2012, Vol. 116, No. 11, pp. 1095–1110

[6] MOL C.D., MOSCI S., *ET AL.*: 'A regularized method for selecting nested groups of relevant genes from microarray data', *J. Comput. Biol.*, 2009, Vol. 16, No. 5, pp. 677–690

[7] YANG J.C., WRIGHT J., HUANG T., MA Y.: 'Image super-resolution via sparse representation', *IEEE Trans. Image Process.*, 2010, Vol. 19, No. 11, pp. 2861–2873

[8] GROSS R., MATTHEWS I., *ET AL.*: 'Multi-PIE', *Image Vis. Comput.*, 2010, Vol. 28, No. 5, pp. 807–813

[9] MA X., HUANG H., WANG S.P., QI A.C.: 'Simple approach to multiview face hallucination', *IEEE Signal Process. Lett.*, 2010, Vol. 17, No. 6, pp. 579–582

Selected content from Electronic Letters

Efficient naturalistic approach to image contrast enhancement

Qiqiang Chen Yi Wan

Institute for Signals and Information Processing, Lanzhou University, People's Republic of China
E-mail: wanyi_js@163.com

Abstract: Contrast enhancement is an important step in many image processing and analysis applications. Although the idea of enhancing image contrast is simple, most published mechanisms of letting computers do so automatically tend to be complex or not so intuitive, and the performance is generally not consistent across different image shooting conditions. A new naturalistic approach to this problem that directly mimics what a human artist would do for contrast enhancement is proposed. The goal is to make every non-noise detail easily perceivable by a normal human. Specifically, the gradient/contrast and local background are computed for each pixel and then linearly mapped to an objective point in the gradient−background plane. Then, a cross-bilateral filter is used to smooth the two linear map parameter images. A moving local window is used to obtain the local noise level for each pixel, which is used to determine whether the pixel should be enhanced. The proposed approach is easy to implement and leads to superior results compared with typical state-of-the-art methods, as confirmed by experimental results.

Introduction

Many images have hidden details because of non-ideal shooting conditions, which adversely affect applications, such as image analysis. Many popular and state-of-the-art methods follow the framework of target function mapping. This mapping *f* can be from the intensity range of [0, 255] onto itself. In [1, 2], *f* is generated through a histogram equalisation or specification. Celik and Tjahjadi [3] obtain *f* by generating a two-dimensional (2D) target image histogram through a Frobenius norm minimisation process. Lee *et al.* [4] do so by emphasising those adjacent pixel pairs whose intensity difference occurs more frequently. The function *f* can also be defined on other relevant domains. In [5], the illumination information is first extracted from the input image. Then, its weighted histogram is converted to a target one, which is taken as a logarithmic curve, through which *f* is obtained on the luminance value. The approach of unsharp masking (e.g. [6]) does not generate such an *f* but uses the unsharp masking filter to essentially perform edge sharpening.

Fig. 1 shows a 1D illustration of edge enhancement. Suppose l_1 and l_3 are fixed. Then, a human artist would typically use the dashed curve to enhance the two jumps at t_1 and t_2. Yet this is not possible or easy to do with most existing methods (e.g. [1, 2, 4]). In the following, we propose a new naturalistic approach to image enhancement, which aims to mimic a human artist by trying to do the essence shown in Fig. 1 and make every non-noise detail easily perceivable to the human eye.

Proposed approach

We first convert an image in RGB space to the hue saturation value (HSV) space and then process the V channel. When a human artist attempts image contrast enhancement by hand, two major considerations are the local background *B* and contrast/gradient *G*. In this Letter, for each pixel (*x*, *y*), we use its local mean value in a 3 × 3 window to define its local background as

$$B(x, y) = \frac{1}{9}\sum_{i=-1}^{1}\sum_{j=-1}^{1} V(x+i, y+j) \tag{1}$$

and we use the simple Sobel operator to compute its gradient (G_x, G_y) and define its scalar gradient value

$$G(x, y) = \sqrt{G_x^2 + G_y^2} \tag{2}$$

To convert the (*B*, *G*) pair to a desired value (B', G'), we propose the following locally linear transform

$$V' = aV + b \tag{3}$$

It is then easy to derive that

$$B' = aB + b \tag{4}$$

$$G' = aG \tag{5}$$

To obtain the desired (B', G') pair, unlike previous approaches we do as a human artist would do: we pick

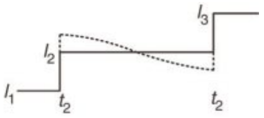


Figure 1 *Illustration of edge enhancement*

With l_1 and l_3 levels fixed, dashed curve is what human artist would typically do to enhance both edges at t_1 and t_2

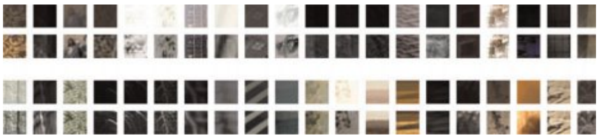


Figure 2 *Forty small image patches of size 20 × 20 (odd rows) with enhanced results through (3) (even rows)*

a and *b* values for each patch were tried out by human volunteers for best perceived result

40 small image patches, see Fig. 2. For each patch, we ask volunteers to adjust the values of *a* and *b* until the 'perceived' best result is obtained. Then we arbitrarily pick two different edge points in the patch and compute their $(B,\ G)$ values and $(B',\ G')$ values before and after enhancement. Thus, we end up with a total of 80 matching pairs of $(B,\ G)$ and $(B',\ G')$. Then we extend these mapping pairs to the following 2D polynomial mapping of order 3 on the $(B,\ G)$ plane:

$$\begin{bmatrix} B' \\ G' \end{bmatrix} = \sum_{i=0}^{3}\sum_{j=0}^{3-i} c_{ij}B^iG^j \tag{6}$$

where the 2 × 1 vector coefficients c_{ij}'s (a total of 20 scalar variables) can be estimated through the well-known least squares minimisation from the 80 matching pairs of $(B,\ G)$ and $(B',\ G')$.

As can be seen from Fig. 3*a* and *b*, the 2D polynomial of order 3 does a reasonable job at approximating the mapping between $(B,\ G)$ and $(B',\ G')$ obtained by human volunteers. Also, it is clear from Fig. 3*c* that for virtually all $(B,\ G)$ values, the human eye tends to favour enhanced results that have a larger value of G, while too dark or too bright values of B tend to become less so.

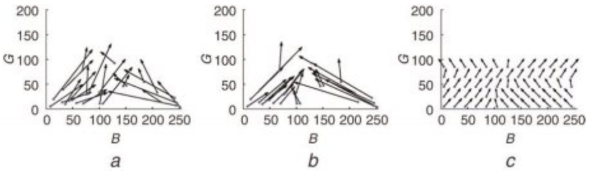


Figure 3 *General mapping between (B, G) and (B′, G′)*

a Some of pairs of $(B,\ G)$ and $(B',\ G')$ obtained from Fig. 2 For each arrow, starting blue point is original $(B,\ G)$ and arrow head is $(B',\ G')$
b Mapping from (6) on same $(B,\ G)$ points as in Fig. 3*a*
c Direction field of mapping from (6) Points above G value of about 100 not plotted since no such training points were available

In practice, one would desire to enhance all image details other than noise. To estimate the noise level L_n, we adopt the well-known wavelet transform method proposed in [7]. Then for any pixel $(x,\ y)$, if $G(x,\ y)$ is sufficiently large or larger than the local noise level, that is

$$G(x,\ y) > \max\{1,\ hL_\mathrm{n}(x,\ y)\} \tag{7}$$

we map the $(B,\ G)$ pair at this point to $(B',\ G')$ according to (6). Otherwise, we deem the point $(x,\ y)$ as a non-edge point and leave it unprocessed ($a = 1,\ b = 0$). The parameter h can be tuned to get varying results. In this Letter, we set $h = 3$.

After the above enhancement on mostly the image edges, we end up with two new images of *a* and *b*. Yet because of discontinuity, they cannot be directly applied to the original image. We next smooth them using the cross-bilateral filter [8], that is

$$a'(x,\ y) = \sum_{(x',y')\in \mathrm{V}(x,y)} w(x,\ y,\ x',\ y')\,a(x',\ y') \tag{8}$$

$$b'(x,\ y) = \sum_{(x',y')\in \mathrm{V}(x,y)} w(x,\ y,\ x',\ y')\,b(x',\ y') \tag{9}$$

where

$$w(x,\ y,\ x',\ y') = \frac{1}{Z(x,\ y)}\mathrm{e}^{-\Big(\Big((x'-x)^2+(y'-y)^2\;\;/2s_s^2\;\;-\Big(\Big((V(x',y')-V(x,y))^2\Big)/2s_v^2\Big)} \tag{10}$$

in which s_s and s_v are empirically chosen to be 10 and 15, respectively, and $Z(x,\ y)$ is the normalising constant so that

$$\sum_{(x',y')\in \mathrm{V}(x,y)} w(x,\ y,\ x',\ y') = 1$$

With the above established key steps, a complete implementation procedure of the proposed approach can be summarised as the algorithm below, steps 0 to 6.

Step 0: Preparation.

(a) For each of the 40 image patches in Fig. 2, convert it from RGB to HSV space, then transform the V values according to (3), where *a* and *b* are picked by human visual inspection to produce enhanced results.

(b) For each of the 40 image patches in Fig. 2, arbitrarily visually pick two edge points. For either of the two edge points, compute its $(B,\ G)$ values according to (1) (2) and its corresponding (B', G') values in the enhanced image patch according to (4) and (5).

(c) Using the 80 pairs of (B, G) and (B', G') from (b), obtain the coefficient values of c_{ij}'s in (6) through the least squares minimisation procedure.

Step 1: Given any RGB image I with a value range of [0, 255], convert it to the HSV space.

Step 2: For each point (x, y) in the V channel image, choose a 16 × 16 neighbourhood centred around it. Apply the Haar wavelet transform on this neighbourhood. Compute the noise level $L_\mathrm{n}(x, y)$ as

$$L_n(x, y) = \frac{\mathrm{Median}\{|u_j|\}}{0.6745} \tag{11}$$

where $\{|u_j|\}$ is the set of the absolute values of all the HH sub-band wavelet transform coefficients in the first level [7].

Step 3: In the V channel, for each pixel (x, y), compute its (B, G) values as in (1) and (2); if $G > \max(1, hL_\mathrm{n}(x, y))$ with $h = 3$, then compute its corresponding (B', G') values as in (6), using the c_{ij}'s obtained in Step 0. Afterwards obtain the a and b values for the pixel (x, y) from (3) and (4); otherwise, set $a = 1$ and $b = 0$. This generates an a image and a b image.

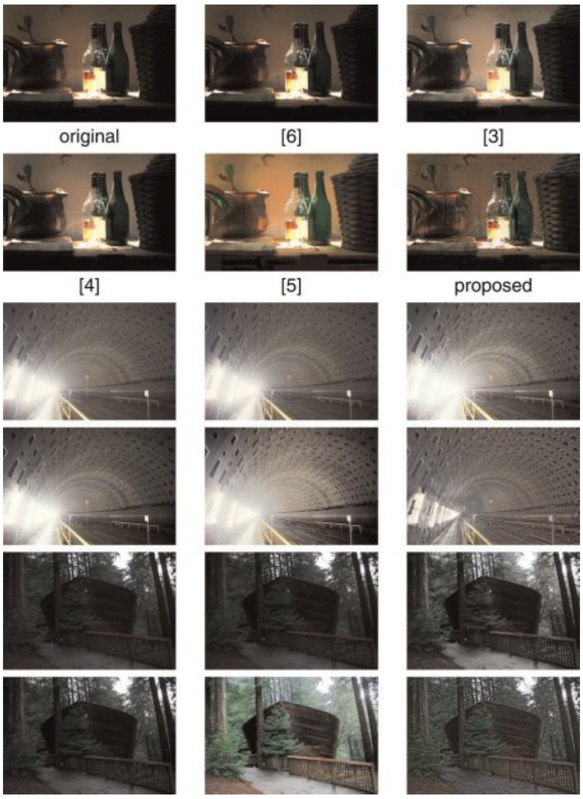


Figure 4 *Image contrast enhancement performance comparison on three test images*

Image positions same for all three cases

Step 4: Apply the cross-bilateral filter on a and b as in (8) and (9) to obtain the new images a' and b'.

Step 5: Obtain the enhanced V' image as

$$V'(x, y) = a'(x, y)V(x, y) + b'(x, y) \tag{12}$$

For those V' values greater than 255 (or less than 0), clamp them to 255 (or 0).

Step 6: Convert the enhanced image HSV$'$ back to the RGB space to produce the final enhanced result.

Experiments

We compared the proposed algorithm with four other representative methods on 30 test images and found that our approach produces the best results. In Fig. 4, we show three of the test images and the enhanced results of different methods. In the first case, the proposed method yields the best overall colour tone and details, especially those on the wall. In the second case of the image tunnel, only the proposed method clearly recovers the row of lights on the left wall and most of the details originally not easily noticeable. In the third case, again more image details can be most comfortably viewed by the proposed method.

Conclusion

In this Letter, we propose a new naturalistic approach to image contrast enhancement. Unlike previous methods, the proposed approach attempts to directly mimic what a human artist would do and enhances every non-noise image feature to a level that can be easily humanly perceivable. This is done by first empirically establishing a mapping on the background–gradient plane, which is then used to linearly enhance the image non-noise edge portion. Afterwards, a cross-bilateral filter is applied to the two linear model parameters so that the entire image can be enhanced with a natural appearance. Experimental results show the advantages of the proposed approach over the state-of-the-art methods.

References

[1] GONZALEZ R.C., WOODS R.E.: 'Digital image processing' (Prentice-Hall, NJ USA, , 2006, 3rd edn.)

[2] WAN Y., SHI D.: 'Joint exact histogram specification and image enhancement through the wavelet transform', *IEEE Trans. Image Process.*, 2007, Vol. 16, No. 9, pp. 2245–2250

[3] CELIK T., TJAHJADI T.: 'Contextual and variational contrast enhancement', *IEEE Trans. Image Process.*, 2011, Vol. 20, No. 12, pp. 3431–3441

[4] LEE C., LEE C., KIM C.S.: 'Contrast enhancement based on layered difference representation of 2D histograms', *IEEE Trans. Image Process.*, 2013, Vol. 22, No. 12, pp. 5372–5384

[5] WANG S., ZHENG J., HU H.M., LI B.: 'Naturalness preserved enhancement algorithm for non-uniform illumination images', *IEEE Trans. Image Process.*, 2013, Vol. 22, No. 9, pp. 3538–3548

[6] POLESEL A., RAMPONI G., MATHEWS V.J.: 'Image enhancement via adaptive unsharp masking', *IEEE Trans. Image Process.*, 2000, Vol. 9, No. 3, pp. 505–510

[7] DONOHO D.L.: 'De-noising by soft-thresholding', *IEEE Trans. Inf. Theory*, 1995, Vol. 41, No. 3, pp. 613–627

[8] LV X., CHEN W., SHEN I.: 'Real-time dehazing for image and video'. Proc. 2010 18th Pacific Conf. on Computer Graphics and Applications, Hangzhou, China, September 2010, pp. 62–69