Volume 8 | ISSN 2041-5923
2016-2017

The Institution of Engineering and Technology

The Best of IET and IBC 2016-2017

INSIDE Papers and articles on electronic media technology from IBC 2016-2017 presented with selected papers from the IET's flagship publication *Electronics Letters.*

Published by the Institution of **Engineering and Technology**

www.**theiet**.org/ibc

IET.tv

ACCESS 8,000+ VIDEOS!

Access Telecommunications content with the world's largest collection of **engineering & technology videos**

Visit www.iet.tv today to;

- Access a huge range of engineering and technology content across 10 specialist channels
- Learn from today's top thought leaders from inspirational events and expert communities
- Participate in live webcasts with prestigious IET presenters
- Stay up-to-date with cutting edge industry information

Tune in to the Communications channel today at:

www.iet.tv

Contents

VOLUME 8 2016

Selected Content from IBC 2016

Selected Content from the IET

Introduction

Welcome to *The Best of IET and IBC* 2016–17. This is the eighth 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 and, 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 display in IBC's Future Zone – an exhibition space where the world's most hi-tech media companies and research organisations proudly demonstrate their very latest concepts and experimental technologies. This year, the immersive technologies of 360° video and VR dominate the Zone. Here, you can not only *see* tomorrow's media but you can personally *experience* it, leaving impressions that will remain with you long after you have left Amsterdam.

We then present eight papers chosen as the best contributions to IBC2016 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, 'WiB – A New System Concept for Digital Terrestrial Television,' and papers representing other hot topics of 2016: HDR (High Dynamic Range) processing and subjective quality evaluation, Object-Based Broadcasting, Virtual Production, 360° Video, Crowd-Based News-Gathering and Gaze-Tracking.

We are also pleased to present personal interviews with individuals whose significant work appears in this volume. First, Erik Stare, Jordi Giménez and Peter Klenner, authors of IBC2016's Best Conference Paper, who discuss their ground-breaking work on digital terrestrial television. Find out how they came to work together as international collaborators from Sweden, Spain, and Germany; where their inventive ideas came from; how they see the future for their new developments; and what the most memorable parts of their project have been.

We then interview IBC's Best Young Professional, Lucien Lenzen from RheinMain University of Applied Sciences, a young researcher whose inventive signal processing work has resulted in a technique for using HDR broadcasts to improve the pictures displayed on standard, non-HDR ('legacy') receivers. Get a glimpse of his personal world and find out what motivates him to work in this area. Also find out what he thinks about the future of immersive media.

From *Electronics Letters* this year we include a selection of media-related papers which have been published since IBC2015. *Electronics Letters* has a very broad scope, covering the whole range of electronics-related research, and the papers chosen this year are those that 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 from more than 320 synopses submitted to us this year by potential authors from across the world. In fact, we were surprised and delighted by this unprecedented number, which is almost 40% up from last year. These synopses provide us with a fascinating barometer of trends in media developments and they continue to show increasing diversification of applications and services, as well as a breathtaking acceleration in technological advances. We have responded to this through the introduction of five new conference session topics this year: IP studio developments, technologies for personalised advertising, 360° TV and VR, new ideas for metadata, and the synchronisation and personalisation of multi-screen media. Our regular 'Cutting Edge Technologies' session will continue to explore the very latest concepts from the labs, and this year we are also giving a special emphasis to our session on assistive technologies for sensory-impaired viewers.

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 chosen 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 watch the IBC website 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 have. 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 IBC2016.

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 Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 1–2

The IET

The IET is one of the world's leading professional societies for the engineering and technology community, with more than 167,000 members in 150 countries and offices in Europe, North America and Asia-Pacific. It is also a publisher whose portfolio includes a suite of 30 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 2015 *Electronics Letters* celebrated their 50th anniversary.

Working closely with the IET Journals team is the IET Sectors team. Sectors work collaboratively with Industry, Academia and Government to engineer solutions for our greatest societal challenges looking at topics such as changing population, future cities, big data and cyber security. Working closely with Sectors, the Communities team exist to act as a natural home for people who share a common interest in a topic area; foster a community feeling of belonging and support dialogue between registrants, the IET and each other. 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.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 1–2

Editorial

A Glimpse of the Future at IBC2016

The Future Zone at IBC provides a great opportunity for visitors to gain 'hands-on' experience of the latest ideas, innovations and inventions in the broadcast and entertainment industries. The Zone is a carefully curated collection of the latest cutting-edge projects and prototypes from leading R&D labs around the World, brought together into a single exhibition area in the Park Foyer, next to Hall 8, in the RAI Centre.

The technologies that are exhibited in the Future Zone come from all types of companies and research groups: from large broadcasting organisations like NHK and the BBC; to the smallest start-ups still in their incubation stage, and university researchers. As long as the projects are interesting and valid, and are not products that are already on sale in the marketplace, we will consider them for inclusion in the Future Zone … choosing the most innovative, the most interactive and sometimes even the most crazy.

The range of this year's Future Zone exhibits is once again highly international, with representatives from Japan, China, Korea, Europe, North America and beyond. The technologies on show broadly divide into three areas:

- studios, production and displays of the future
- virtual and augmented reality applications
- posters and 'blue sky' research topics

We have incorporated two special improvements into the Future Zone layout this year, to enhance the visitor experience:

- a new 'digital display' area for IBC Conference Posters. The Posters are the highly respected, rigorously peer-reviewed project descriptions and ideas, chosen by the IBC Technical Papers Committee for their relevance to the themes of the IBC Conference Programme and their capacity to intrigue. This year, the Posters will be displayed on large interactive LED display screens, enabling the Poster authors and audience to interact more effectively with the powerpoint slide shows, videos, and images about their individual projects.
- a 'Future Realities' Theatre, where there will be a programme of short, informative talks by industry leaders on the advances in technologies and practical applications in the virtual and augmented reality (VR and AR) sectors. In this high intensity programme of presentations, we will learn about the market drivers, business cases and potential pitfalls for VR and AR in broadcasting and entertainment. There is no need to book your place for these talks, just turn up and participate.

The whole Future Zone is free and open to all, thoughout the IBC Show; and we draw your attention in particular to the highly-popular IET Reception, to be held in the Zone's Future Realities Theatre at 15:00 on Friday the 9th September. This event features a keynote opening address by Naomi Climer, President of the IET and Chairman of IBC Council, who will talk about both the technology advances on show in the Zone, and improving diversity for the next generation of engineers in our industry. There will be complimentary refreshments at this event, which will also herald the start of the 'Future Realities' programme of presentations.

We cordially invite you to visit the Future Zone during the Show at IBC, immerse yourself in the hands-on experiences and get a 'real' glimpse of the future … it is truly the jewel in the crown of the IBC Show.

PROF. DAVID CRAWFORD
IBC Conference Executive Producer
University of Essex, and Ravensbourne

WiB – a new system concept for digital terrestrial television (DTT)

E. Stare[1], J.J. Giménez[2], P. Klenner[3]

[1]*Teracom, Sweden (erik.stare@teracom.se)*
[2]*Universitat Politècnica de València, Spain (jorgigan@iteam.upv.es)*
[3]*Panasonic Europe Ltd, Germany (peter.klenner@eu.panasonic.com)*

Abstract: A new system concept for DTT, called "WiB", is presented, where potentially all frequencies within the Ultra High Frequency (UHF) band are used on all transmitter (TX) sites (i.e. reuse-1). The interference, especially from neighbouring transmitters operating on the same frequency while transmitting different information, is handled by a combination of a robust transmission mode, directional discrimination of the receiving antenna and interference cancellation methods. With this approach, DTT may be transmitted as a single wideband signal, covering potentially the entire UHF band, from a single wideband transmitter via the TX site. Thanks to a higher spectrum, utilisation, the approach allows for a dramatic reduction in fundamental power/cost and approximately a 37-60% capacity increase for the same coverage as with current DTT. High speed mobile reception as well as fine granularity local services would also be supported, without any loss of capacity. The paper also outlines further possible developments of WiB, e.g. doubling the capacity via cross-polar Multiple In Multiple Out (MIMO), backward-compatible with existing receiving antennas, and adding a *second*, WiB-mobile, Layer Division Multiplexing (LDM) layer within the same spectrum, either as a mobile broadcast or as a mobile broadband.

Introduction

Basic principles of WiB

WiB is a new wideband reuse-1 based DTT concept, developed at Teracom, Sweden, by E. Stare. WiB builds on the earlier work of *Cloud Transmission* by Wu et al. (1), and is radically different from conventional DTT and offers very attractive characteristics. In a traditional High Power High Tower (HPHT) DTT Multi-Frequency Network (MFN) or Single Frequency Network (SFN), a high capacity is typically transmitted per UHF channel, e.g. 33-40 Megabits per second (Mbps) with the Digital Video Broadcasting (DVB)-T2 standard (2). However, the high order modulation that is needed to carry the high capacity makes the signal sensitive to interference, which requires transmitters that operate on the same frequency to be positioned sufficiently far away and in a regular pattern. This way, their respective signals are attenuated so as not to cause harmful interference when they are received. When SFNs are used, the same principle of separation applies to groups of SFN transmitters. A consequence of this approach is therefore that only a fraction (1:N) of the frequencies at a particular site are actually used, which is called reuse-N frequency planning (for DTT, N is typically in the range 4 to 7). Unfortunately, since required power, according to Shannon (3), fundamentally increases *exponentially* with capacity, high-power transmitters (TXs) are necessary.

With WiB, a far more power-efficient approach is employed, which is to use potentially *all* UHF channels from all TX sites (reuse-1) and spread out the transmitted power equally across these channels. This can be achieved potentially as a single wideband signal using a single TX, where the existing 0.2-0.4 MHz (2.5-5%) spectral gaps between UHF channels could also be exploited. Using reuse-1 and e.g. Quadrature Phase Shift Keying (QPSK) code rate at half the modulation allows for mobility and a spectral efficiency of about 1 bps/Hz. This is about the same as with an "all DVB-T2" implementation of DTT using the existing type of frequency planning/reuse (assuming 5×40 Mbps MFN or 6×33 Mbps SFN). In both cases (taking overhead into account), approximately 200 Mbps can be offered within the 224 MHz of the DTT spectrum (470-694 MHz) that remains after the 700 MHz-band release. Simulations, however, indicate that WiB could be used with significantly higher spectral efficiency than 1 bps/Hz (1.37 to 1.60; see chapter about performance results below). The use of a very robust mode may also eliminate the need to use a Guard Interval (GI) altogether, since at low C/N levels, the gain of using a GI seems to be lower than the overhead "pain".

A commonly employed DVB-T2 mode is 256-Quadrature Amplitude Modulation (QAM) code rate 2/3. However, with a QPSK rate of half the required TX power (for a given coverage) is about 50 times (17 dB) lower than the 8 MHz channel. The net effect of this is that a WiB signal would fundamentally only require about 10% of the *total* TX power (of all multiplexes) for DVB-T2, see Figure 1. There are, however, also other WiB gains that may further reduce the required power, see below.

Peak service data rate and tuner bandwidth

Since the basic coding/modulation is restricted with WiB, due to the reuse-1, the capacity within a *single* UHF channel will be limited, in the order of 7-10 Mbps. To compensate for the lower capacity within a particular UHF channel, it is instead assumed that a service can be spread across *several* UHF channels. Assuming the basic tuner bandwidth is increased by, for example, a factor of four, i.e. from one UHF channel (8 MHz) to four UHF channels (32 MHz), this would allow the peak data rate to be increased by a factor of four to around 28-40 Mbps within this wider bandwidth. A side effect of the increased tuner bandwidth is an increased frequency diversity, which generally improves performance, assuming the service is appropriately interleaved across the entire bandwidth. This effect can be maximised by interleaving a service across the total used spectrum, e.g. by using frequency-hopping techniques like Time-Frequency Slicing (TFS), see (2) and Giménez et al. (4).

Interference considerations

With reuse-1, the receiver will experience a much lower Carrier-to-Interference ratio (C/I) than is usual for DTT and this issue must, of course, be seriously considered. The first tool to handle interference is the robust mode (e.g. 17 dB is more robust than current DTT), which may allow a C/I close to 0 dB. The

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 4–9

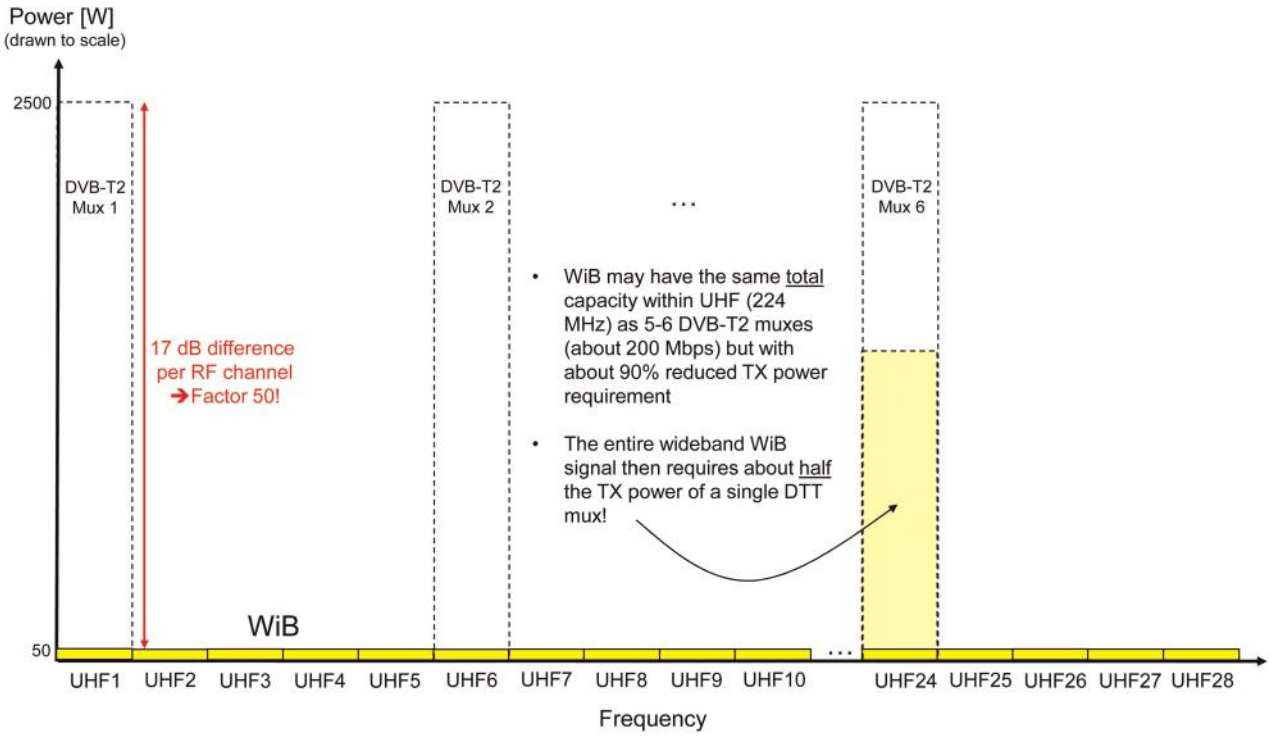


Figure 1 *Required power for DVB-T2 and WiB*

second tool (for fixed roof-top reception) is that a directional antenna typically offers a very significant (up to 16 dB) *discrimination*, i.e. "attenuation" of signals in unwanted directions or polarisations. Finally, there are methods for interference cancellation, by which unwanted signals may be cancelled under certain conditions (see Interference Cancellation chapter below). It should be noted that thanks to the inherent ability to cope with interference from adjacent transmitters, the WiB concept allows potentially all TXs to transmit different content.

Network cost savings with WiB

Savings in capital expenditures (CAPEX)

Perhaps the most striking (CAPEX) cost saving is due to the fact that the total required TX power of the *equipment* could fundamentally be reduced by about 90%, thereby considerably simplifying the infrastructure. This could allow all existing TXs to be replaced by a single wideband TX with a lower power (about 50%) than *each* of the traditional DTT TXs. Some performance requirements, such as linearity, of the (single) TX could also be greatly relaxed thanks to the robust signal, which could simplify the design of the TX and also contribute to higher power efficiency. Furthermore, since the complete WiB signal can be transmitted as a single wideband signal, there is no need to use RF combiners anymore – there could be just a single exciter and a single wideband TX, with a Radio Frequency (RF) filter. A 90% reduced power also reduces the cooling requirements very significantly and allows for simple battery back-up power solutions for many smaller sites.

Due to the lower power/cooling/volume/weight of the overall TX equipment, it could more easily be installed in the TX mast, which would also eliminate the need for RF feeders. WiB also lends itself well to be used together with "active TX antennas", i.e. a group of antenna elements (or even each antenna element) could potentially have its own very low power, wideband "mini-TX", which could enable an electronically-controlled *phased array antenna*, whereby the antenna diagrams could be tailor-made and optimised to the desired characteristics.

Savings in operational expenditures (OPEX)

Similar to the CAPEX case, the reduced *power consumption* would be the most striking OPEX advantage, but in this case, the cost reduction is seen in a reduced electricity bill. In addition to the fundamental 90% power reduction, there are also other possible factors that could allow for a further reduction, in consumed power, such as elimination of the attenuation in (now superfluous) combiners, feeder, RF split etc., which may amount to a total of about 3 dB and a consequent further power reduction of 50%. On top of that, increased frequency diversity may offer further power consumption gains, due to a better link budget. Depending on the type of TX, there may also be power efficiency gains in the actual TX implementation due to lower linearity requirements. From a service perspective, the overall complexity of the system would be reduced since there are fewer system components and the sensitivity of these is also reduced thanks to a far more robust operation mode. Lower power may also generally increase transmitter lifetime and reduce failure probability. Furthermore, there will be a reduced or no need for frequency changes or frequency re-planning once the network is in operation.

Interference cancellation with WiB

Layer division multiplexing-based interference cancellation (LDM-IC)

When a target signal S1 is interfered by an equally-modulated but stronger S2 signal, LDM-IC (also referred to as Successive Interference Cancellation) can be implemented by first demodulating the stronger signal, then re-modulating it and subtracting it from the incoming signal, which is possible since it is perfectly known after the (assumed successful) demodulation. In a final step, the target S1 signal can then be demodulated. This process is possible as long as the actual Carrier-to-Noise +Interference ratio (C/(N+I)) of the signal to be demodulated (here S2 followed by S1) is larger than the required C/(N+I). For a 1 bps/Hz spectral efficiency (e.g. half the QPSK rate), the required C/(N+I) is close to 0 dB. The described process can be generalised

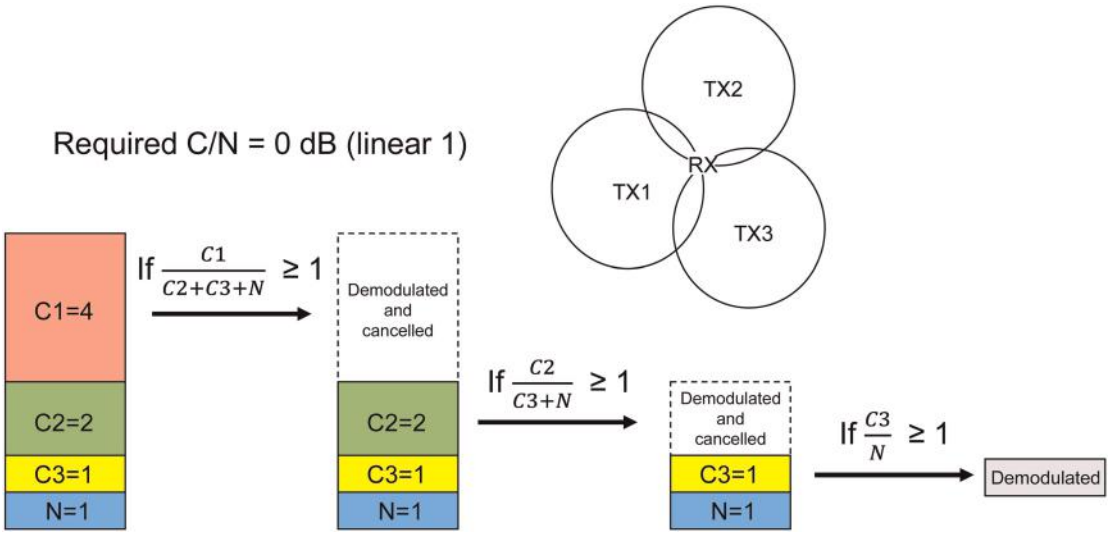


Figure 2 *Example of LDM-IC, where the weakest TX3 signal (received with power C3) is demodulated in three steps*

and used for any number of signals as long as the C/(N+I) requirement is fulfilled for every demodulated and subtracted signal.

One example scenario is shown in Figure 2 above, where the received power from TX1, TX2 and TX3 are C1, C2 and C3 respectively, with N being noise power. The weakest TX3 signal can be demodulated as long as the C/(N+I) is fulfilled in the cancellation process of the two stronger signals and the required C/N is fulfilled for TX3.

For LDM-IC to work (with reasonable complexity) in WiB, all involved transmitted signals need to be fully synchronised with aligned Forward Error Correction (FEC) blocks, i.e. the reception situation needs to look similar to a traditional SFN, but with the different TXs transmitting *different* content. To allow for LDM-IC, each received TX also needs to include scattered pilots that are orthogonal to all other received TX signals involved in the interference cancellation, i.e. there needs to be at least three orthogonal pilot patterns.

Antenna-based interference cancellation (ANT-IC)

A completely different approach to interference cancellation is to use multiple receiving antennas, e.g. arranged as a phased-array antenna. Even in the simplest case, with two dipoles, these could have an electronically-controlled beam, which could dynamically maximise the C/(N+I) of the signal to be received, such as by full cancellation of one signal.

Combination of LDM-IC and ANT-IC

The most powerful, approach for interference cancellation identified to date is to combine LDM-IC with ANT-IC in such a way that for each signal to be demodulated, the C/(N+I) is maximised by appropriately adjusting/optimising the electronically-controlled antenna for this particular signal. When this signal has been demodulated/cancelled, the antenna may be retuned and the C/(N+I) can again be maximised for the next signal to be demodulated.

Variable bit rate (STATMUXED) services using multiple PLPs

Variable bit rate services may be transmitted over DVB-T2 using variable bit rate Physical Layer Pipes (PLPs). However, with WiB LDM-IC would not generally work with variable bit rate PLPs, since the signals from the associated TXs (varying independently) would then not be synchronised as required. With WiB, all PLPs have an equal and constant bit rate instead, and the VBR aspect is catered for by dynamically mapping a VBR service to a variable number of PLPs; the number being dependent on the instantaneously required bit rate of the service. The efficiency of the statistical multiplexing would not be significantly affected by this.

Performance results from simulations

Spectral efficiency limits with HPHT network modelling

The achievable spectral efficiency of WiB in a HPHT network was estimated via Monte Carlo simulations. The network was modelled by a homogeneous reuse-1 hexagon lattice, see Figure 3, with 60 km distance between adjacent TXs and with an effective antenna height of 250 m. Spectral efficiency was evaluated at the assumed worst point of the network, i.e. the mid-point between three

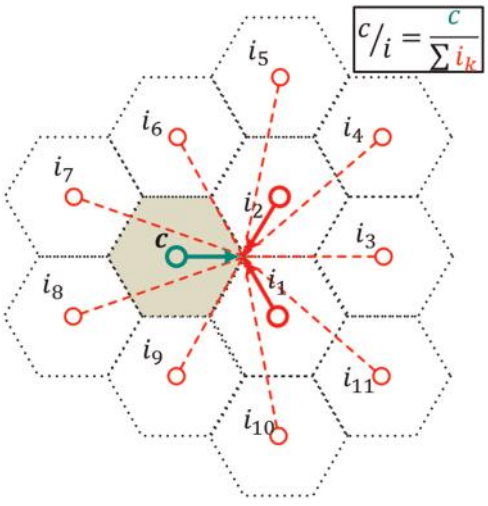


Figure 3 *Hexagon lattice*

Table 1 WiB spectral efficiency

Time correlation type	Best TX	Wanted TX
Inter/Intra site (C)	3.41 bps/Hz	1.55 bps/Hz
Intra-site (U1)	3.38 bps/Hz	1.37 bps/Hz
No correlation (U2)	4.07 bps/Hz	1.60 bps/Hz

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 4–9

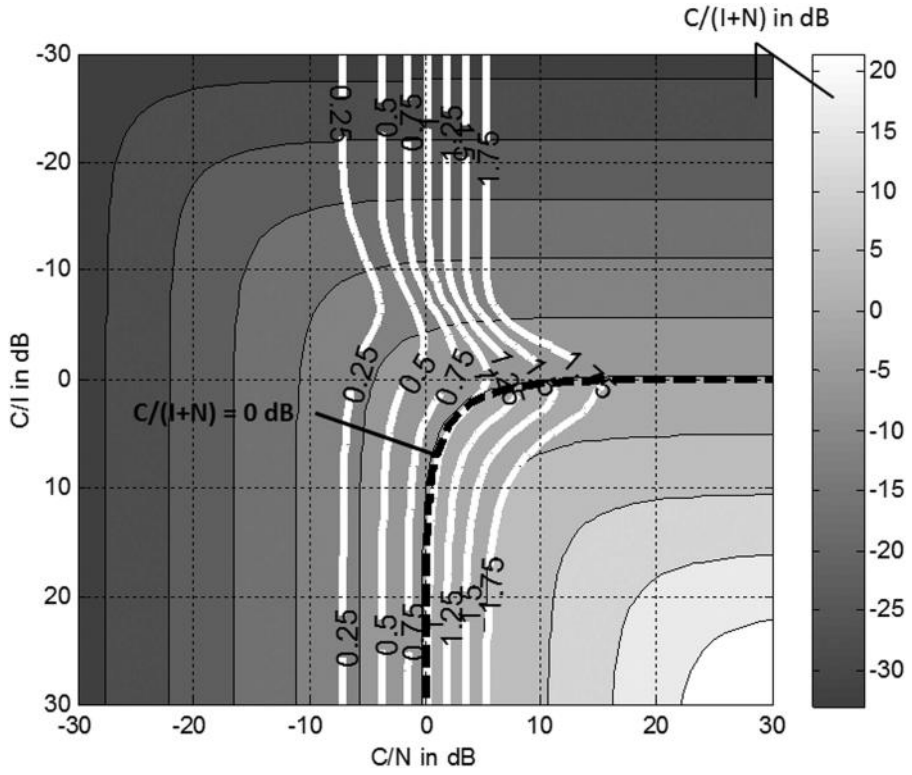


Figure 4 *Achievable rates 0.25…1.75 bps/Hz for a two-TX system with an applied random phase and optimal detection*

adjacent TXs, at 10 m above ground level, see Figure 3. The assumed Effective Radiated Power (ERP) was 1 kW per 8 MHz UHF channel and the receiving antenna gain was 11 dBd, with discrimination according to International Telecommunication Union (ITU) Rec. BT.419 (5). The assumed down-lead loss and receiver noise figure were 4 dB and 6 dB respectively. The propagation model was according to ITU Recommendation ITU-R P.1546 (6) over land. Time variations of the propagation was statistically modelled by fitting two log-normal distributions to the propagation curves given for 50% and 10%, and 10% and 1% of time, respectively. This allowed three different correlation models to be used: full inter-site as well as intra-site correlation (C), no inter-site but full intra-site correlation (U1), no inter- or intra-site correlation (U2). Fading was statistically modelled by means of two processes. A frequency-independent but location-dependent fading was modelled as a log-normal random variable with 0 dB mean and 5.5 dB standard deviation. A site-to-site correlation model is applied to account for location-dependent fading on angular position and distance between stations (7). A frequency dependent fading process with 2 dB standard deviation was added to model potential frequency-dependent variations of the received field strength, according to Giménez et al. (8). The coverage criterion was to maintain a reception with 95% location probability and for 99% of the time at the worst point in the network. For each realisation, the C/(N+I) determined the maximum Shannon capacity that can be transmitted. In order to account for ideal frequency interleaving, the average spectral efficiency over all RF channels was calculated.

Two application cases were considered. *Best TX* models a receiver pointing to the best TX in each realisation. *Wanted TX* models a receiver pointing to a desired (not necessarily the best) TX among the three closest transmitters. The spectral efficiency that the WiB system can provide is calculated so that layers above the desired one can be cancelled. Since all layers provide the same capacity, the minimum of all of them is selected. Table 1 shows the achievable spectral efficiency with 95% location probability for 99% of time.

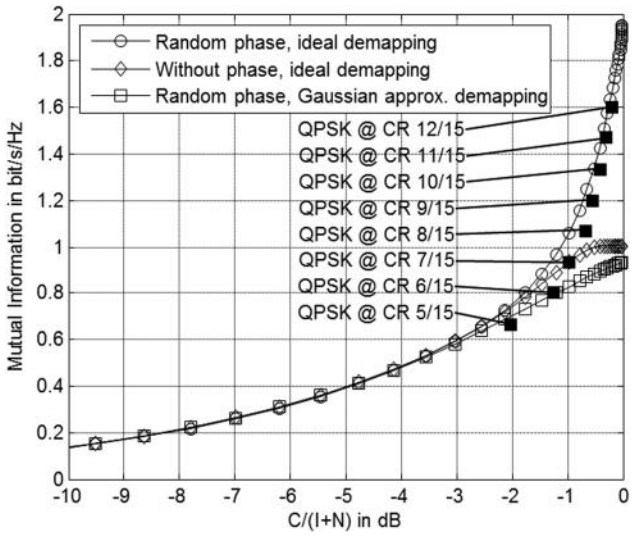


Figure 5 *For C/I=0 dB, achievable rates for various de-mappers. Included is the performance with ATSC 3.0 CRs (■)*

Optimised performance for QPSK interfered by QPSK

In the HPHT network simulations above, all interfering signals that are weaker than the one currently being demodulated have been treated as noise. However, this is a *pessimistic* assumption since the constellation is known, and this *a priori* information can be exploited to improve performance. From an information-theoretical point of view, based on the concept of Mutual Information, one can derive the theoretically optimum performance for a general case where a QPSK signal suffers interfered by another QPSK signal, having an added random phase. The random phase will usually arise naturally as a result of different path delays in the network combined with interleaving, but can be added intentionally at the transmitter to avoid a destructive superposition

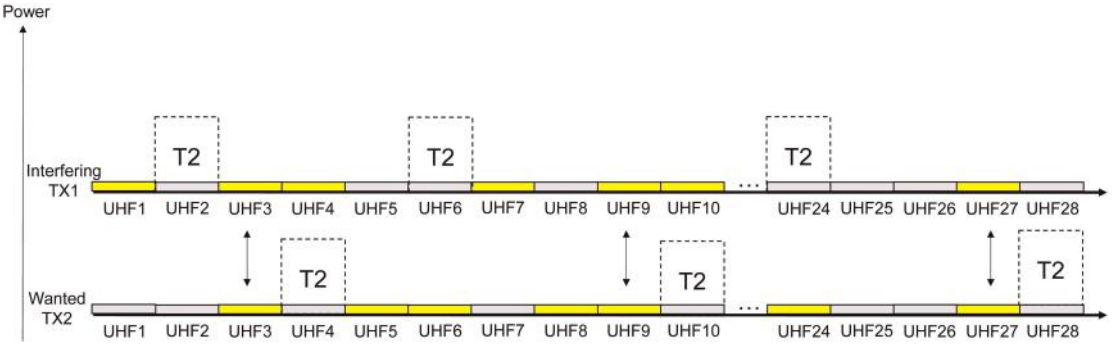


Figure 6 *WiB interleaved with DTT*

of cells at the receiver under highly correlated Line-of-Sight (LoS) propagation conditions. This has been done and the required C/N for a given power relation of the two signals (C/I) is given in Figure 4 for a number of different spectral efficiencies (bps/Hz). For 1 bps/Hz, with required C/N=0 dB, adding a noise-like interferer at C/I=0 dB would leave no room for noise thus causing the required C/N to approach infinity. However, when the QPSK constellation is taken into account in an optimum way (ideal demapping) with 2-dimensional log-likelihood ratios (2D-LLRs), the required C/N becomes 6.0 dB. Exploiting the knowledge of the constellation therefore seems to allow very large gains in performance. Network performance is expected to be further improved when this behaviour is exploited. Figure 5 depicts the achievable rates for a fixed C/I = 0 dB assuming optimal and suboptimal de-mapping. Depicted are also rates for Advanced Television Systems Committee (ATSC) 3.0 (9) with QPSK and different code rates (CRs) based on the presence of a random phase, optimal de-mapping and sum-product FEC decoding. As can be seen, ATSC codes closely follow the ideal curve.

Receiver aspects

The receiver could basically be a single-tuner Orthogonal Frequency Division Multiplex (OFDM) receiver (or other Multi-Carrier system) with at least a 32 MHz tuner bandwidth. The wider *virtual* bandwidth (up to e.g. 224 MHz) is either handled by TFS (frequency hopping) or by sampling at a higher rate and then performing the frequency hopping in the digital domain. To maintain the same OFDM symbol time and carrier spacing in this 32 MHz bandwidth as with the DVB-T2 32K mode in 8 MHz, a Fast Fourier Transform (FFT) size of 2^{17} (128K) would be required. For channel estimation, three different channel estimates (one for each received TX signal in a 3-TX environment) need to be handled and passed through the (cell-based) de-interleaving chain, so that the LDM-IC process can be performed after a single "once-and-for-all" de-interleaving step. The complexity of the fundamental FEC-decoding itself could be similar to DVB-T2 or ATSC 3.0 (or maybe lower, because of the lower expected peak bit rate), but the interference cancellation would require the basic FEC decoding to be perfomed at a higher rate (or with more parallelism).

Considering that Moore's Law has been in action since the specification and implementation of DVB-T2 in 2009 and is expected to be in operation until the implementation of WiB sometime in the 2020's, the overall WiB receiver complexity does not appear, in fact, to be overly complex.

Introduction scenarios

All migrations to new standards are more or less difficult and can normally not be achieved in "one shot". Instead, a gradual approach is typically preferred. Two different approaches have been identified to introduce WiB: "*Dedicated band*" or "*Interleaved*". With the dedicated-band approach, part of the UHF band would be released for WiB in a similar way as the 700 MHz band will be released, with similar international co-ordination. A contiguous sub-part of the UHF band could then be used for WiB, ideally with a synchronised introduction. In later steps, this band could potentially be extended to the full UHF band. With the interleaved approach, WiB would be introduced in much the same way as DTT was introduced in an analogue TV context, i.e. current DTT would still be used and WiB would be transmitted with *some* power from potentially all frequencies that are not used by current DTT. The transmitted WiB power would be adjusted for each TX site/frequency individually, so as not to cause harmful interference. In a transition phase, it might be necessary to accept some degree of degradation of the existing DTT, although this may also be compensated for by using a somewhat more robust transmission mode for DTT. It is possible to design WiB to allow the use of an arbitrary subset of UHF channels (with the particular subset varying across TX sites), still allowing for interference cancellation, see Figure 6. This subset could first be small and then grow until WiB eventually might replace DVB-T/T2 altogether.

Further work, advanced WiB

The basic WiB system outlined earlier in this paper could be further developed in a number of different ways. Here we will only mention two such directions, both of which could double the WiB capacity: WiB-MIMO and WiB-LDM. With WiB-MIMO, one would add a second independent signal on the opposite polarisation and effectively have a dual-Single Input Single Output (SISO) signal, where one of the signals would be receivable by a legacy single-polarisation TV antenna, thanks to antenna polarisation discrimination. By frequency-transposing one of the two received polarised components immediately after the antenna, the existing down-lead could be used for both components and the receiver could be an SISO receiver, which would choose a selected service from the relevant component. With WiB-LDM, one would add a second WiB-mobile, LDM layer superimposed on the WiB-DTT signal, and transmit this from the TXs of a densified network providing mobile coverage of the WiB-Mobile signal. The weaker WiB-DTT signal would still be strong enough for rooftop reception after cancellation of the stronger WiB-Mobile signal. In one "fully converged" scenario, the WiB-Mobile signal could even be a mobile broadband (bi-directional unicast) signal, e.g. as part of *5G New Radio*.

Conclusions

This paper has presented a new DTT system concept called WiB, based on Wideband reuse-1, which allows for a potentially very large reduction of DTT network costs (both CAPEX and OPEX) while significantly exceeding the capacity and coverage that can be obtained with an optimised implementation of DVB-T2, for example. In the most demanding "Wanted TX" case, simulations

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 4–9

indicate a possible capacity increase in the range of 37-60% assuming the DVB-T2 reference can carry about 200 Mbps per site within the 470-694 MHz band. The complexity of the receiver could be limited to that required to receive a particular service and not the entire transmitted WiB capacity. Furthermore, the WiB concept also allows for high-speed mobile reception, without the need for handover or capacity loss, fine content granularity (no big SFN areas required) and a possible reduction (or even elimination) in GI overhead. Finally, WiB can be extended in various ways, such as backward-compatible cross-polar MIMO, dual-layer LDM mobile/fixed reception and even with a WiB-DTT and WiB-mobile broadband sharing literally the same spectrum.

Acknowledgments

The authors would like to thank Oliver Haffenden (BBC) for valuable comments. Erik Stare would also like to thank Magnus Ahxner and Staffan Bergsmark (both at Teracom) for valuable discussions and inspiration.

References

1 Wu, Y, Rong, B, Salehian, K, Gagnon, G.: 'Cloud transmission: a new spectrum-reuse friendly digital terrestrial broadcasting transmission system', *IEEE Trans. Broadcast., September* 2012, **58**, (3), pp. 329–337
2 ETSI EN 302 755: 'Digital video broadcasting (DVB); frame structure channel coding and modulation for a second generation digital terrestrial broadcasting system (DVB-T2)'
3 Shannon, C.A.: 'A mathematical theory of communication', *Bell Syst. Tech. J.*, July, October, 1948, **27**, pp. 379–423, 623–656
4 Giménez, J, Stare, E, Bergsmark, S, Gómez-Barquero, D.: 'Time frequency slicing for future digital terrestrial broadcasting networks', *IEEE Trans. Broadcast.*, June 2014, **60**, (2), pp. 227–238
5 ITU Recommendation ITU-R BT.419-3: 'Directivity and polarization discrimination of antennas in the reception of television broadcasting', 1992
6 ITU Recommendation ITU-R P.1546-5: 'Method for point-to-area predictions for terrestrial services in the frequency range 30 MHz to 3000 MHz', 2013
7 Saunders, S, Zavala, A.A.: 'Antennas and propagation for wireless communication systems' (Wiley, New York, NY, USA, 2007, 2nd edn.)
8 Giménez, J, Gozálvez, D, Gómez-Barquero, D, Cardona, A.: 'A statistical model of the signal strength imbalance between RF channels in a DTT network', *Electron. Lett.,* 7th June 2012, **48**, (12)
9 ATSC 3.0 Candidate Standard: 'Physical Layer Protocol (A/322)', available at: www.atsc.org

Interview - Erik Stare, Jordi Giménez, Peter Klenner

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

ES: I am a Senior R&D Engineer at the Swedish broadcast network operator Teracom. I have taken part in the development and standardisation of DTT since the very beginning in the early 90's and have actively contributed to the standardisation of DVB-T, DVB-H, DVB-T2, DVB-NGH and most recently ATSC 3.0. I have also contributed to some WorldDAB standardisation, e.g. DAB+. My current work focuses on two areas: next generation broadcast systems (physical layer) and integration of DTT into home (WiFi) networks.

JG: I work as an R&D Engineer at the ITEAM Institute of the Universitat Politècnica de València (UPV), Spain. I have been involved in the research of next generation terrestrial broadcasting since I joined iTEAM in 2011, with particular attention to the standardisation of DVB-T2/NGH and ATSC 3.0 technologies. Such a close link with industry allowed me to earn a PhD in Telecommunications from UPV (2015) and to conduct several research stays at Teracom (Sweden, 2013), Wireless@KTH (Sweden, 2014) and Institut für Rundfunktechnik (Germany, 2015).

PK: I am a Research Engineer at Panasonic Europe in Frankfurt, Germany. I joined Panasonic in 2011 after graduating with a PhD in Communications Technology. After a short detour in the standardisation activity of DVB's stereoscopic video, I now work on Digital Terrestrial Television with a focus on standardisation.

2. You are all from different parts of Europe. How did you find yourselves working together?

Combined answer:

Teracom, UPV and Panasonic have all co-operated in DVB standardisation and Study Mission activities for a number of years, and each of us have been a part of this. Most recently we have also co-operated in ATSC 3.0 standardisation. As a spin-off of these standardisation activities Teracom and UPV have also had bilateral co-operation regarding DTT network-related studies, especially related to Time Frequency Slicing (TFS).

3. Have your diverse backgrounds contributed to the success of your project?

Combined answer:

Yes, definitely. Our diverse backgrounds allowed us to focus in-depth on different aspects of the WiB system. Erik's large experience in terrestrial broadcasting together with LDM, PHY-layer simulation (Peter) and background from TFS, network simulation and planning (Jordi).

4. DTT is over 20 years old and most of us thought that it was as good as it could get. What inspired you to seek a more efficient approach?

Combined answer:

We were of course inspired by earlier work on *Cloud Transmission* performed by CRC in Canada, in which a robust mode was combined with reuse-1. The DVB Study Mission work on a Next Generation Terrestrial (NGT) system followed by ATSC 3.0 standardisation work both triggered us to look into areas of improvement.

One source of inspiration was the desire to search for the best possible DTT system, when network/frequency planning and cost are taken into account. We had previously observed, in connection with TFS network studies, that there was a strong tendency for improved overall spectral efficiency when the frequency reuse factor was reduced below traditional values.

The problem of using a higher proportion of frequencies, or even all frequencies, from a given site is of course the increased interference (reduced C/I) that would result. However, with interference cancellation methods, which is a new element into the equation, such interference can to a large extent be removed, provided a robust transmission mode is used. Interference cancellation is very effective for the level of robustness we foresee for WiB, but not very useful in a traditional DTT context.

Another source of inspiration was the observation that for a given total capacity per site (within the total available spectrum) it is fundamentally more power-efficient to use a higher proportion of all frequencies (per site), together with a robust transmission mode, rather than a lower proportion of frequencies with a higher capacity per frequency. The most power-efficient scheme is to use all frequencies on all sites, like we do with WiB.

For the wideband aspect of WiB we were inspired by earlier studies by ourselves (ES, JG) on TFS, which have demonstrated the large performance gains that a higher (virtual) bandwidth can offer.

Finally, one could also say that the "environment" is changing (less spectrum given for DTT and higher quality expectations from consumers) and therefore a change may also be required in technology. This has also inspired us.

5. Should non-specialists reading your paper be concerned that your results are based on idealised lattices and statistics, rather than on real-world terrain?

Combined answer:

Idealisations in science help us to think and, in this particular case, to predict the upper limit of a WiB system's performance. It should be noted that both network-level-type and system-level-type simulations have been performed.

To be able to predict how well WiB would work in a particular network and terrain, the use of real-world terrain data, transmitter site positions etc would of course be beneficial. However, one would then only know about the WiB performance for the particular studied area. The results could then be said to be too specific to the particular conditions of that area and that other areas would allow better or worse performance. It would thus be difficult to extrapolate the results to other areas or to have more generally-valid results.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 10–11

The other approach, that we have used and which is very common, is to have a generic model that tries to model a typical (idealised) network and to see how well the system performs in that model network. The network planning modelling is however based on huge amounts of measurements from many countries. It is therefore believed that the results we have obtained should be relatively representative of what performance could be achieved in a real network, although there could be deviations in particular areas. It should be noted that although some aspects of the simulation results are idealised in an optimistic direction there are also several aspects that can be said to be pessimistic, i.e. where a real-world situation could allow even better performance.

Of course the kind of study we have performed could and should also be followed by more studies using real-world terrain, to confirm our results.

6. Most of the world already operates DTT networks, won't this limit the impact that your new ideas can have?

Combined answer:

The question seems to presuppose that DTT does not need evolution and will remain as it is. However, the current spectrum trends, together with increased service quality (HDTV, 4K for DTT) and the limitations of DVB-T2 to enable them, reveal that something should change if DTT wants to remain competitive.

WiB could be used in HPHT networks as a dedicated Next Generation Terrestrial system, where increased capacity and far lower costs would be the driving forces. The longer the expected lifetime is for dedicated DTT services the more rationale there should be for this approach. It should be noted that change of terrestrial standard has already occurred twice (analogue to DVB-T, DVB-T to DVB-T2), so it is not a new concept.

In another scenario WiB principles are integrated into future 3GPP standards and implemented on an LPLT, HPHT or hybrid HPHT/LPLT infrastructure. For 5G New Radio one might even imagine broadcast and mobile telecom unicast using the same spectrum, separated by interference cancellation methods. Such a system could result in huge benefits both for the traditional broadcast world and for the mobile telecom world.

Finally, the far lower power requirement of WiB is not only a cost advantage but could be considered to be a very important "green" technology advantage, which may be increasingly appreciated in the future.

7. Was there a 'Eureka moment' or specific breakthrough that determined the success of your project?

Combined answer:

The conceptual WiB idea did not occur suddenly "in the bathtub" but evolved more gradually, based on some ideas that were floating around and gradually were combined and integrated into a coherent WiB concept. In order of appearance, the advantages of using a wideband system (TFS) came first, followed by the realization of the importance of using lower frequency reuse. The full appreciation of the potential of interference cancellation allowed this to stop at reuse-1. When the consequences of reuse-1 were analysed it became apparent that this could bring very significant cost advantages.

One Eureka moment was the realisation that cross-polar MIMO could be introduced in a way which would be backward compatible with legacy single-polarisation receiving antennas. Since cross-polar MIMO has always been seen as incompatible with legacy receiving antennas this came as a huge surprise.

Another significant moment occurred when we saw from simulations that the WiB system with >1bps/Hz was actually possible, and also when we saw that the non-Gaussian behaviour of the interference could even provide larger efficiencies in reality.

It was also very satisfying when the simulation results showed that the bit-error performance with real ATSC-LDPC codes and QPSK interference were close to the information-theoretical limits within the expectable margin.

8. Your paper mentions several options for future development of your ideas. Which do you consider most promising?

Combined answer:

The backwards-compatible cross-polar MIMO option may be the easiest one since it could directly reuse the basic WiB system in a dual-SISO way and would (normally) allow reception of one polarisation with legacy antennas.

However, the option with the highest potential (but maybe with greatest uncertainty) is probably the "5G New Radio" variant where broadcast and mobile telecom would use the same spectrum and be separated by interference cancellation. This would allow a converged world of mobile communication and broadcast for the provision of media and entertainment services at anytime, anywhere and to any device.

9. Tell us about your personal preferences in entertainment media; are you, for example, multi-screen users, downloaders or live big-screen sports viewers?

ES: Most of the time I use a big screen for viewing, very often with my wife. What I watch is typically drama series, films and various types of non-fiction (e.g. news, documentary, science). The majority of the time I use pre-recorded (via PVR) material or Netflix/HBO-type streaming.

JG: I am not a strong entertainment media consumer but am enthusiastic about media technology from traditional DTT, satellite and shortwave listening, to VoD, streaming or HbbTV. I have really appreciated the advances in video and sound quality in recent years, which really make you feel part of a film, a concert or a sports event.

PK: I like the dialectic of the US-serial format and, if the family schedule permits it, am a bit of a marathon-viewing fan. For this purpose, I love today's streaming services on a big-screen.

HDR for Legacy Displays Using Sectional Tone Mapping

L. Lenzen

RheinMain University of Applied Sciences, Germany

Abstract: High dynamic range (HDR) allows us to capture an enormous range of luminance values within a still image or a sequence of video frames. But many consumers will not have the necessary displays to experience this in the near future. To allow these 'legacy' users to benefit, an adaptation using global tone mapping would be a possible solution. But the results suffer from low subjective contrast and can product large-area flicker. To overcome these drawbacks, three enhancement steps are proposed. They are based on certain broadcast requirements and on viewer preferences, surveyed at the beginning of this study. The basic idea is to analyse each luminance value for its relevance in the image and discard unimportant ones. This 'virtual aperture' will be processed across the whole image and on image sections. Finally, the tone mapping result will be composed with the chrominance values by using a modified IPT colour space.

Introduction

One of the main goals of future television is to create a more immersive experience. The viewer should get the feeling he or she is inside the action. One key component to achieve this is HDR. By preserving details in highlights and shadows, simultaneously, a visual sensation is created close to viewing the real scene. For efficient coding, Hybrid-Log-Gamma and Perceptual Quantizer (PQ) are discussed. This technology will produce a significant increase in subjective video quality as shown in several studies (e.g. 1, 2). But for the next few years, we can expect over 90% of viewers will still have legacy displays. So one main question that comes with HDR is: 'Is there a way to let the viewers with legacy displays also benefit from HDR?' (Note that a faithful reproduction of the images in this paper is only available when viewed as a .pdf on a sRGB display.)

Requirements and Goal

When we talk about showing HDR material on a standard dynamic range (SDR) display, we have to think about how to handle the enormous dynamic range at capture, compared to the small dynamic range of a particular display. The easiest way is to cut off all parts outside the range, but clipping would bring us back to the burned-out highlights and the undefined shadows. A more promising way is to do a contrast compression also known as tone mapping. Our goal is to create a tone mapping system that is capable of live broadcast and does not suffer from the typical problems of tone mapping as discussed later.

At first, we define the following requirements for live broadcast:

- The most important point is to create a very pleasing image that matches the viewer's preference, although a very pleasing image is not equivalent to the most realistic and natural image. In addition, the look should not be too different from the familiar look of television today, to gain the viewer's acceptance.
- The system must be very stable, so it can be used for live broadcast with no post production. Never should tone mapping artefacts like flicker, ghosting, or halos, be produced.
- Unlike an aperture, which has discrete steps, a smooth adjustment to changing brightness during a scene should be performed. This transition should not take place if there is a scene change; in this case, the exposure ought to adapt to the new situation immediately.
- To utilize the capabilities of legacy displays, the system should be sensitive to the display and environmental brightness and should adapt to these parameters, depending on the scene.
- For live broadcast, it, is necessary for the system to work in real-time.

Related Work and Evalutation

Before engineering, we ran a test, letting 5 professional colourists grade 7 HDR sequences[1] for SDR displays. In a viewing session, 20 subjects (mostly non-expert viewers) would decide which grading he or she liked best. Every grading was compared one-to-one with every other, so the subjects only had to decide which one of the two was more pleasing. The experiment showed the viewers liked high contrast and high saturation. This was confirmed in several previous studies (3,4,5). But sometimes, excessive saturation was found, and high saturation does not lead to a pleasing image. It seemed rather important that all information within the image could be easily extracted. The viewer likes to see all the details in all parts of the image, without having to pay too much attention. Saturation and contrast often enhance this aspect. Clipping is acceptable, in this case, and was sometimes missed when it was not there. For example, to reproduce spectral highlights on SDR, clipping is an adequate process. Another outcome is that the white point is only of secondary importance. A warm grading was not preferred over a cold one, as one might expect, so the white point is only a creative decision and need not be adjusted by the tone mapping.

Discovering a viewer's preference in tone mapping is not an easy task. Tone mapping operators (TMOs) are often categorised into local and global. Local operators analyze the neighbourhood of each pixel. They create a spatial region of adaptation. That is why they can preserve small local contrast. On the other hand, this is the reason they tend to produce visual artefacts, such as: halo, ghosting, or flicker. In addition, the real-time requirement does not make it easier to use a local tone mapping. Some of the results appear a little surreal and are far away from the familiar TV look. Also agreeing with the results of Petit and Mantiuk (3),

[1]*bistro_01, bistro_02, carousel_fireworks_01, carousel_fireworks_08, carousel_fireworks_09, cars_fullshot and fireplace_01* filmed by Fröhlich et. al (7)

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 12–17

Eilertsen *et al.* (5) and Aydin *et al.* (6), local operators do not seem to be the first choice for live broadcast.

The second category consists of global operators. They create one transfer curve for the entire image (like a camera transfer curve), mostly considering the brightest pixel and average luminance. Usually, this is less computationally intensive. Moreover, global operators are more stable. They can produce flicker when only based on an intra-frame analysis, but smoothing by interpolation can be easily performed and could reduce, or even eliminate, flicker. However, the key issue of global tone mapping is the images suffer from low subjective contrast - they look like log-signals. Not surprisingly, the transfer curve is close to a log-curve.

Our idea is to improve subjective contrast, saturation and sharpness by combining classic global tone mapping with three enhancement steps, which we call "EVI" (Enhanced Video Information).

Enhancement Steps

Step 1: Virtual Aperture

A log-curve has a flat gradation. Therefore a higher dynamic range can be coded with the same number of bits. You need a bigger change in luminance to produce the same change in coded values. If you display such a log-curve directly, the image looks like someone has overlain a grey veil. Contrast and 'crispness' are missing. The larger the difference between the brightest and darkest pixel, the flatter the gradation created by the global TMO. But is there really a need for such high dynamic range in every image? Which information is necessary and which can be seen on a SDR display?

For example, let us look at the sequence filmed by Fröhlich *et al.* (7). The image is picked from the sequence, called 'bistro' (Figure 10 first image - at the end of the paper). The highest luminance value is about 58253 cd/m^2 and the lowest about 0.16 cd/m^2 (with noise). So, we are talking about a dynamic range of about 18.5 stops. If we remove only one per mill on each side, the luminance varies between 6653 cd/m^2 and 0.24 cd/m^2. It is reduced by nearly 4 stops. This is a prime example, but we can conclude that, often, extreme values, which are unimportant for the image (sometimes only noise), influence the result of tone mapping. As a result, a flat gradation is applied. Beyond that extreme values have high deviations. These cause the typical large-area flicker, because brightness varies from one frame to another. So it is worth considering how to handle the extreme values.

Some TMOs perform clipping by cutting-off the highest and lowest percent. But such fixed values would not lead to the best results, because not every scene needs a steep gradation. For some images, the impact of 1% might be subjectively less and, for others, subjectively more. That is why our system should be smarter when adapting automatically to the dynamic range distribution within a scene. A possible means of doing so would be to judge the importance of a luminance value, based upon a histogram.

The luminance Y_{in} of the input is calculated, e.g., by using the form mentioned below for Rec.709 in linear light. The calculation differs if another camera or input gamut is used.

$$Y_{in} = 0.2126 \cdot R_{709} + 0.7152 \cdot G_{709} + 0.0722 \cdot B_{709}$$

After that, the histogram $h(Y_{log})$ over all pixels of the incoming image is computed (Figure 1). The histogram has about 13 orders of dynamic range (Figures 1 to 5 show a section of 8 orders). For the y-axis, the luminance Y_{in} is scaled in a logarithmic way:

$$Y_{log}(Y_{in}) = \begin{cases} 0, & \text{for } Y_{in} \le 10^{-k_1} \\ \log_{10}(Y_{in}) + k_1, & \text{for } Y_{in} > 10^{-k_1} \end{cases}$$

k_1 should be chosen, so that no negative values are produced. As a next step, we introduce a so-called 'contrast box'. The contrast box has a constant dynamic range independent from its position on the y-axis, due to the log-scale. The width W is defined, dependent upon the highest display luminance L_{dmax} (measured in cd/m^2), because the brighter the display, the more dynamic range can be preserved, without creating a flat gradation by using tone mapping. k_2 and k_3 are fixed scaling factors.

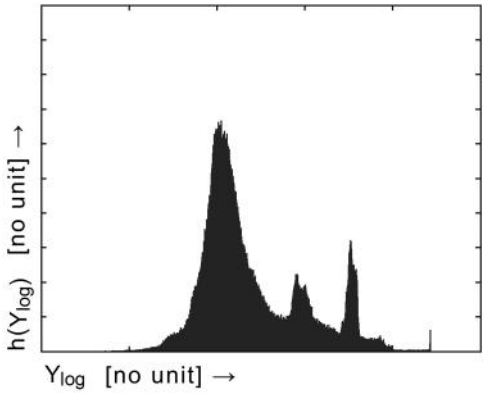


Figure 1 *histogram $h(Y_{log})$ of the incoming image*

$$W = k_2 + k_3 \cdot L_{dmax}$$

This box is moved to the position where the most luminance values are inside the box (Figure 2). The region inside the box could be interpreted as the centre of interest. Here, we would like to have no clipping, but a steep gradation. Let us call the middle of the box μ (on the x-axis). It is useful to compute the cumulative histogram $H(Y_{log})$ and search for the position g_l with the highest slope at the distance W.

$$H(Y_{log}) = \sum_{k=0}^{Y_{log}} h_k$$

$$H(Y_{log} + W) - H(Y_{log}) = \max \rightarrow g_l$$

$$\mu = g_l + 1/2 \cdot W$$

The importance of the centre could be linked to the number of values inside the box. More values mean more importance. To normalize the number, a division by the width W is performed. This leads to

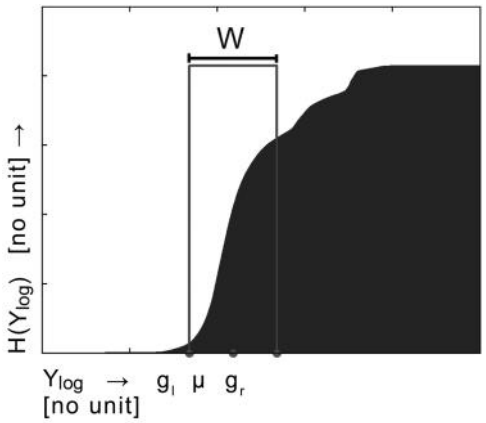


Figure 2 *cumulative histogram $H(Y_{log})$ with contrast box*

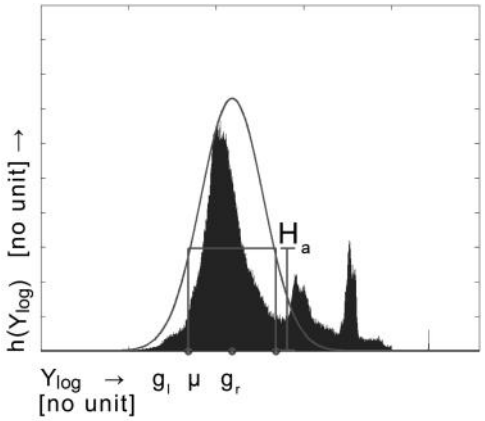


Figure 3 *histogram $h(Y_{log})$ with contrast box of height H_a and Gaussian function*

the height H_a of the contrast box (Figure 3).

$$H_a = \frac{\sum_{g_l}^{g_l+W} h_k}{W} = \frac{H(g_l + W) - H(g_l)}{W}$$

When we define a standard deviation σ linked to H_a, we can calculate a Gaussian function of the form

$$\sigma = \left(\frac{H_{av}}{H_a}\right)^{k_5} \cdot H_{av} \cdot k_6$$

$$g(Y_{log}) = \frac{H_a \cdot k_4 \cdot W}{\sqrt{(2\pi \cdot \sigma^2)}} \cdot e^{\left(\frac{-(Y_{log}-\mu)^2}{2\cdot\sigma^2}\right)}$$

dependent on the dynamic range distribution (Figure 3). k_4, k_5 and k_6 are fixed scaling factors for the height and width of the Gaussian function. H_{av} is an average height of the contrast box. With a very centred dynamic range distribution, a high Gaussian function with very steep slopes is formed. When the values are highly distributed over a large dynamic range, the slopes become flat. The function $g(Y_{log})$ is used to weight the original histogram h (Y_{log}) by multiplication.

$$f(Y_{log}) = h(Y_{log}) \cdot g(Y_{log})$$

This leads to an overshooting around the centre of interest (Figure 4). By comparing $f(Y_{log})$ to the clipping thresholds for the shadows and highlights ($s_1 \cdot L_{room}$, $s_2 \cdot L_{room}$; where L_{room} is the environmental luminance), the centre would not lose information. The comparison leads to the darkest and brightest logarithmic luminance value,

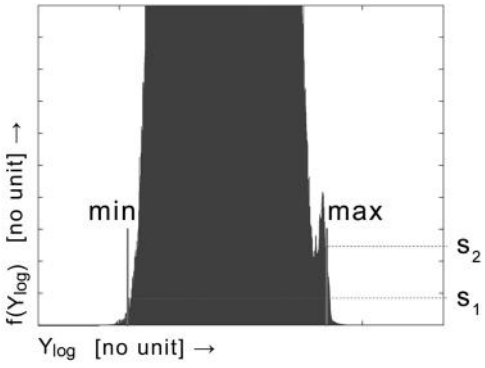


Figure 4 *$f(Y_{log})$ with overshooting around the center of interest*

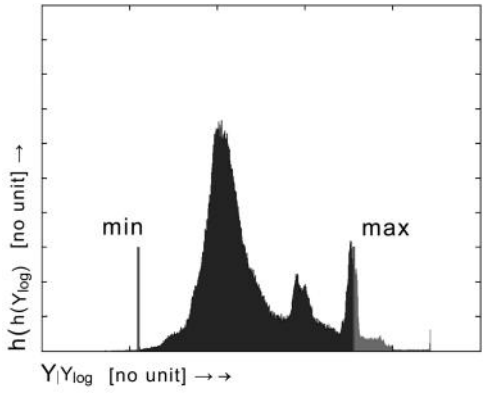


Figure 5 *histogram $h(Y_{log})$ with clipping thresholds*

which is important for the image (Figure 5). Afterwards they can be transformed backwards into luminance values (Y_{max}, Y_{min}).

By avoiding extreme values, large-area flicker is reduced remarkably, but it has not been resolved. μ and σ can vary from one frame to another, too. Using an interpolation of the form

$$\mu_{gn} = \mu_{g(n-1)} \cdot z_\mu + \mu_n \cdot (1 - z_\mu)$$

$$\sigma_{gn} = \sigma_{g(n-1)} \cdot z_\sigma + \sigma_n \cdot (1 - z_\sigma)$$

helps to improve this. z_μ and z_σ have values between 0 and 1. The default value is 0.95. In this case the new μ and σ would have 5% impact on the overall μ (μ_{gn}) and overall σ (σ_{gn}), respectively. Such a transition should not take place if there is a scene change. Here, the exposure must adapt to the new situation immediately. Therefore, a scene change detector is needed. The scene detection should be coupled to the centre of interest and, therefore, to μ and σ. If one of the following conditions is not fulfilled, a new scene is assumed. s_μ and s_σ are the thresholds.

$$\left|\frac{\mu_n - \mu_{g(n-1)}}{\mu_{g(n-1)}}\right| \le s_\mu \left|\frac{\sigma_n - \sigma_{g(n-1)}}{\sigma_{g(n-1)}}\right| \le s_\sigma$$

Figure 6 shows a comparison with and without the virtual aperture. For global tone mapping, the Drago operator (8) is used.

Step 2: Sectional Tone Mapping

Global tone mapping has some big advantages, compared with local tone mapping when considering (live) broadcast. But for scenes with a high dynamic range, even the results obtained with the virtual aperture have problems at small local contrast. To overcome this, the virtual aperture must be linked with local adaptation. For this purpose the image is separated into blocks. In experiments, 16×9 led to the best results. The virtual aperture is now processed on every block separately. If we took more blocks, there would not be enough luminance values to compute a good histogram for the virtual aperture. If we took less blocks, the local impact is small. We call this 'sectional tone mapping'. Compared with classic local tone mapping, we can avoid major drawbacks.

To overcome visible steps at the borders of the boxes (Figure 7 middle), their results for Y_{max} and Y_{min} have to be smoothed with those of their neighbours. Therefore, a non-linear function is used. Y_{max} and Y_{min} are matrices of the size 16×9 and not scalars, as with classic global tone mapping. Afterwards, they are resized to the size of the whole image. The matrices will be passed over to the tone mapping operator (e.g. the modified Drago). Y_{in} and Y_{max} are scaled by a so-called 'world adapting luminance' (L_{wa}) as used, e.g., at Drago *et al.* (8), called Y_{win} and Y_{wmax} in the following

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 12–17

Figure 6 *Left: A typical live broadcast situation shown with a camera transfer curve. Middle: The same situation shown with the Drago global tone mapping operator. Right: Using the virtual aperture with the Drago operator.*

Figure 7 *Left: Virtual aperture. Middle: Virtual aperture performed on 9×16 blocks. Right: Sectional tone mapping with z=0.25.*

equitation. L_{wa} can also be a matrix.

$$Y_{out}(x, y) = \frac{1}{\log_{10}(Y_{wmax}(x, y) + 1)} \cdot \frac{\log(Y_{win}(x, y) + 1)}{\log\left(2 + \left(\left(\frac{Y_{win}(x, y)}{Y_{wmax}(x, y)}\right)^{\frac{\log(b)}{\log(0.5)}}\right) \cdot 8\right)}$$

b is set to 0.85 as in (8). For a smooth and stable Y_{max} it is recommended to blend the results of the global and the sectional virtual aperture. z is a value between 0 and 1.

$$Y_{max} = Y_{max\_global} \cdot z + Y_{max\_sectional} \cdot (1 - z)$$

Step 3: Modified IPT Colour Space

When considering tone mapping, we perform a contrast compression of the luminance component, only. So the question: How should we apply these results to the chromaticity, too? The goal is not to influence the hue and saturation. Conventionally, the image would have been converted to the XYZ colour space and tone mapping would be applied on the Y component. After that, Y_{in} would be compared with Y_{out} and the ratio would also be applied to X and Z. This approach effects the colour impression.

For this reason, our system is taking an easy, but powerful, approach by using the IPT colour space – an improvement of the CIELab colour space. The IPT is based on a decorrelation of luminance and chrominance and is perceptually uniform (9). 'I' stands for the intensity and 'P' and 'T' are colour-difference signals (called protan and tritan). So in our approach the incoming signal is always transformed to IPT space.

As usual, the tone mapping gets applied to the luminance component ($Y_{in} \rightarrow Y_{out}$). By using step 2, Y_{out} is automatically normalized to the same range as 'I' (0 to 1). Y_{in} has to be normalized too, so the ratio of tone mapping can be easily applied on 'I'. If this is done, the image looks too dark, overall. This is caused by the fact that the LMS components (long-medium-short cone response data) were raised to the power of 0.43 before being transformed to IPT. So some gamma correction must be applied to the ratio, too.

$$I_{out} = I_{in} \cdot \left(\frac{Y_{out}}{Y_{in}}\right)^{0.43}$$

At the end, the image is transformed backwards to RGB, for example according to Rec.709. Values outside the gamut get clipped. For display, a gamma must to be applied.

When using the IPT colour space as described above, there is a big loss of saturation. The problem is that, although chrominance is separated from luminance, the incoming luminance influences P and T. For higher luminance values, P and T increase - namely at a factor of 2.6915 with every decade. In SDR images, this would not have such an impact, but tone mapping manipulates the luminance very strongly. The luminance after tone mapping does

Figure 8 *The left image shows the tone mapping in XYZ, the right one in IPT. The grass and the sky in XYZ have a little hue shift and look more burned-out.*

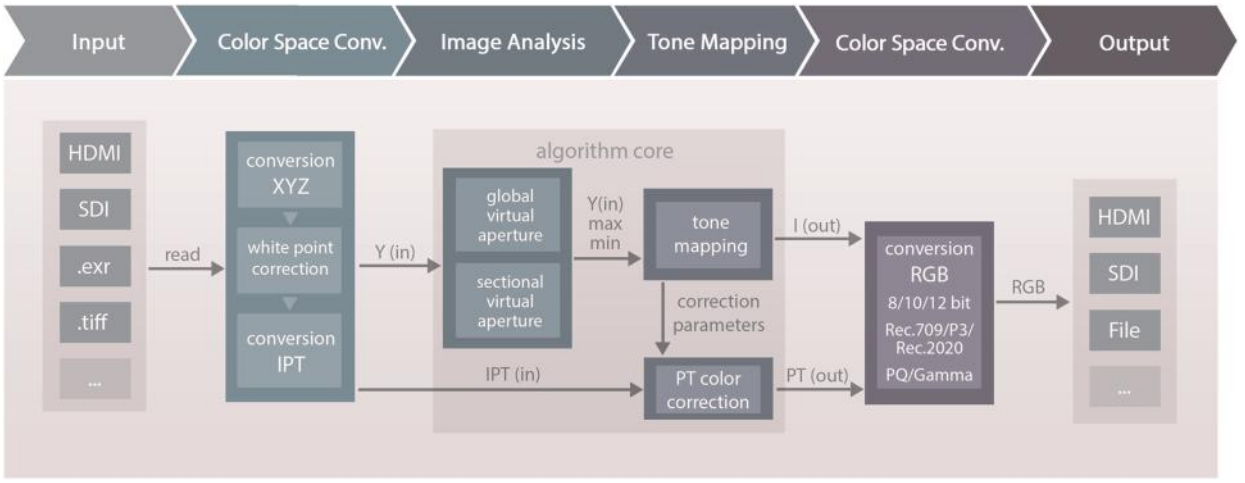


Figure 9 *The incoming HDR images are converted to XYZ and, afterwards, to IPT. The Y component is tone mapped, using the three enhancement steps. P and T are corrected before composing these values with Y_{out} to a new IPT image. Finally, a transformation to the output colour space is performed.*

not correspond to P and T any longer. That is why we use a compensation, inspired by the traditional one in XYZ. However, a full compensation leads to over-saturated results, which is why we added the factor 0.85 (based on subjective tests. T is treated equivalently:

$$P_{out} = P_{in} \cdot \left(0.15 + 0.85 \cdot 2.6915^{\log_{10}\left(\frac{Y_{out}}{Y_{in}}\right)}\right)$$

$$\text{or } P_{out} = P_{in} \cdot \left(0.15 + 0.85 \cdot \frac{I_{out}}{I_{in}}\right)$$

Compared with XYZ, IPT can provide a more realistic saturation, contrast, and colour fidelity. In a direct comparison, as in Figure 8, XYZ looks washed-out. Using IPT also brings back structure in the highlights. In Figure 9, the whole workflow is shown.

Transmission

There are some possible applications for the proposed system in the production and transmission path. E.g., it could be implemented directly in the camera, so a SDR-compatible signal (SDR with EVI) is at the output, and everything else would remain the same. When using it for today's HD transmission, fixed L_{dmax} and L_{room} have to be chosen. 200 cd/m^2 and 10 cd/m^2 for display and environmental brightness, respectively, are a good compromise, which will match most viewing situations. For HDR transmissions in the future, e.g., two fixed pairs of values could be used for creating a two-layer solution, such as the Dolby Version method (10) or Mantiuk *et al.* (11). By generating metadata from the virtual mastering display, the home display can adapt the signal for its specification. For OTT, perfectly adjusted video could be generated at the server and sent out to the home display. Therefore, the home display passes over its L_{dmax} and L_{room}. Last, the algorithm could be implemented in the display - or for legacy displays, in the Set Top Box, directly. So the automatic adaption can be done at home.

Figure 10 *Example images: Left: TV as it is today. Right: EVI*

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 12–17

Conclusion and Outlook

Global tone mapping has much more potential than is exploited in today's workflows. When putting it into an appropriate environment, it can give a large improvement to broadcast pictures, even today. Our system processes HDR content to SDR, without excessively influencing the look of the image. The impression of contrast, saturation, and sharpness is close to today's television picture, so it will be accepted by the viewers, but much more scene contrast is delivered. Two example images are shown in Figure 10. The viewer can easily follow the details in the shadows and is not dazzled by the highlights. The system is stable for all tested material, so far, and fulfils all requirements for live broadcast. A real-time implementation is under development. A further improvement is that the aperture need not be controlled. Also, the system is future-proof. Different display and environmental luminance conditions are considered, as well as wide colour gamut. Beside the real-time implementation, further research will focus on a more complete viewer preference model and investigate different colour spaces and colour corrections for tone mapping to improve the enhancement system.

Acknowledgements

The author would like to thank his colleagues for their contributions to this work and DFL/Sky for the scenes captured in the Allianz Arena. The work was financed by the RheinMain University of Applied Sciences and the Federal Ministry for Economic Affairs and Energy Germany. TransMIT supports the patent process.

References

1 Hoffmann, H.: 'Investigation in an image dynamic range methodology', *NAB Show.*, April 2014.
2 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. 2010, pp. 17–24.
3 Petit, J., Mantiuk, R.: 'Assessment of video tone-mapping: are cameras' S-shaped tone-curves good enough?', *J. Visual Communication and Image Representation.*, 2013, **23**, pp. 1020–1030.
4 Akyüz, A.O., Fleming, R., Riecke, B.E., Reinhard, E., Bülthoff, H.H.: 'Do HDR displays support LDR content? A psychophysical evaluation', *ACM Trans. Graph.*, July 2007, articleno 38.
5 Eilertsen, G., Wanat, R., Mantiuk, R.K., Unger, J.: 'Evaluation of tone mapping operators for HDR-video', *Comput. Graph. Forum.*, 2013, **32**, pp. 275–284.
6 Aydin, T.O., Stefanoski, N., Croci, S., Gross, M., Smolic, A.: 'Temporally coherent local tone mapping of HDR video', *ACM Trans. Graph.*, 2014, **33**, pp. 196:1–196:13.
7 Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., Brendel, H.: 'Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays', *Proc. SPIE.*, **9023**.
8 Drago, F., Myszkowski, K., Annen, T., Chiba, N: 'Adaptive logarithmic mapping for displaying high contrast scenes', *Computer Graphics Forum.*, 2003, **22**, pp. 419–426.
9 Ebner, F., Fairchild, M.D.: 'Development and testing of a colour space with improved hue uniformity', *Proc. IS&T 6th Colour Imaging Conference.* 1998, pp. 8–13.
10 Dolby: 2014. Dolby vision white paper. http://www.dolby.com/us/en/technologies/dolby-vision/dolby-vision-white-paper.pdf
11 Mantiuk, R., Efremov, A., Myszkowski, K., Seidel, H.-P.: 'Backward compatible high dynamic range MPEG video compression', *ACM SIGGRAPH*, 2006, pp. 713–723.

Interview - Lucien Lenzen

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

I graduated with a Master's degree in media and communications technology from the University of Applied Sciences Wiesbaden in 2014. Since then, I am a research assistant and a Ph.D. student.

I am investigating how we can transform the broad range of information captured by an HDR camera so that it can also be displayed on other monitor types than HDR ones. That is important because there is a high diversity at the playback side today. The days are past when everybody had a 100 CD/m2 CRT.

At the moment, I am performing viewing tests with Probands to measure the preferred dynamic range depending on the display capabilities. I want to know how much contrast compression is accepted before the images start to appear in 2D.

In my free time, I do athletics. I also love to go to film festivals – especially for short films because of their high variety – and I am a hobby photographer.

2. What motivated you to take a master's degree in media and communications technology?

On the one hand, there is a love for high-quality images influenced by my hobbies mentioned before. On the contrary, I think media and communications technology is a very exciting field. There are a lot of changes taking place at the moment. I would love to witness this development or even more to be part of it. Especially HDR got quite far in the last years, but there is still enough to be discovered.

3. Your paper explores how legacy displays can make use of HDR transmissions but won't most people just go and buy new HDR displays?

It is not only on the screen. The idea is to have an SDR transmission, but using an HDR camera. Not every broadcaster will have the ability to set up a complete HDR workflow. With the algorithm, they can improve their image quality without changing anything but the camera. Moreover, some displays will not be HDR in the next years. I am thinking about smartphones and Co.

4. Tell us about the most challenging problem you had to solve when developing your algorithm.

Broadcast deals with a very wide range of different scenes. But for Live, we need a system which works automatically and therefore, almost exclusively, fixed parameter values. It leads to an optimization problem which will never be solved completely. I spent a lot of time to find the best adjustments by using as much test material as I could get. But in January 2015 when I started working, there were only few video sequences available. So we had to film some on our own e.g. in the Allianz Arena.

5. Your work involves collaboration with industrial partners, how important has this been for you?

I work at a university of applied sciences, and I think the name speaks for itself: I do not want to do theoretical research which cannot be used for any application. With the help of industrial partners, I can combine my theoretical academic research with a benefit for potential users by providing the opportunities to manufacture my system and providing equipment for testing.

6. What other media research would you like to undertake as part of your Ph.D.?

Classic Tonemapping tries to model the human visual system (HVS). But the goal of television is to create a pleasing image. The result can be very different. As described in the paper, we did the first tests to learn more about the viewer preference, but a whole model is missing. So my intention is to start to set up a "subjective viewer preference model" for HDR down conversion analyzing different scenes and external conditions (display and environmental brightness).

7. What are your ultimate career ambitions?

I have not specified my ultimate goals yet. I As I said in the beginning, there are a lot of exciting developments going on in video technology, and hope I will be able to use them to make the images better and better. Therefore, I believe that industrial research is the most important field. But I realized that I enjoy working with students at the university, too. So I stay open-minded and curious what the future will bring.

8. Tell us about your personal preferences in entertainment media; are you, for example, a multi-screen user, a Downloader or live big-screen sports viewer?

I must confess that most of the time, I am an ordinary TV viewer. It is probably not in spite, but particularly because I work with new technologies the whole day. But of course I also do video streaming for watching series, and I like the big-screen experience at the European or World Sports Championships.

9. HDR will be a big part of future Immersive Media. What do you think will be the most exciting application of Immersive Media?

I think that no single technology can provide a truly immersive experience. It is necessary to have a content-dependent combination of different immersive technologies. What I would love to experience is that the walls of our room consist of displays and that you could change their appearance to whatever you like to see, e.g. full-length windows providing a panorama of skyscrapers at night or small windows in a wooden house with the view of the mountains. Consequently, I would combine a very high spatial resolution with HDR and perhaps with augmented reality.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, p. 18

Gaze tracking using corneal images captured by a single high-sensitivity camera

L. El Hafi, M. Ding, J. Takamatsu, T. Ogasawara

Robotics Laboratory, Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan

Abstract: This paper introduces a method to estimate gaze direction using images of the eye captured by a single high-sensitivity camera. The purpose is to develop wearable devices that enable intuitive eye-based interactions and applications. Indeed, camera-based solutions, as opposed to commercially available infrared-based ones, allow wearable devices to not only obtain natural user responses from eye movements, but also scene images reflected on the cornea, without the need for additional sensors. The proposed method relies on a model approach to evaluate the gaze direction and does not require a frontal camera to capture scene information, making it more socially acceptable if embedded in a glasses-shaped device. Moreover, recent development in high-sensitivity camera sensors allows us to consider the proposed method even in low-light condition. Finally, experimental results using a prototype wearable device demonstrate the potential of the proposed method solely based on cornea images captured from a single camera.

Introduction

Today's high-resolution, high-sensitivity cameras alongside powerful image processing algorithms encourage many new applications. In particular, the increase in resolution and the deceased size of camera sensors allow new eye-tracking methods previously judged impractical.

Moreover, the industry's recent growing interest in virtual reality (VR), augmented reality (AR) and smart wearable devices has created a new momentum for eye tracking. Indeed, eye tracking can be used as an intuitive AR input, or used to reduce motion sickness induced by ill-calibrated VR devices (1). Eye movements in particular are viewed as a way to obtain natural user responses from wearable devices alongside gaze information to analyze interests and behaviors (2).

In this paper, we introduce a method to estimate the gaze direction using cornea images captured by a single high-sensitivity camera. Corneal imaging was first explored in (3) and further refined in (4), (5) and (6). Camera-based solutions, as opposed to commercially available infrared-based (IR) ones, allow wearable devices not only to obtain natural user responses from eye movements, but also scene images reflected on the cornea without the need for additional sensors. In particular, our method does not require a frontal camera to capture the scene, making it more socially acceptable as part of a wearable device.

We use a model-based approach to estimate the gaze direction in our proposed method. First, we reconstruct a 3D eye model from an image of the eye by fitting an ellipse on the colored iris area. Then we continuously track the gaze direction by rotating the model to simulate projections of the iris area for different eye poses and matching the iris area of the subsequent images with the corresponding projections obtained from the model. From an additional one-time calibration step, we can also compute the reflected point of regard on the cornea, enabling us to identify where a user is looking in the scene image reflected on the cornea.

In order to validate our method, we conducted several experiments using different hardware, such as a high-sensitivity camera in low-light condition and glasses equipped with a near-4K camera. We did this in front of a computer display to demonstrate the potential of such an eye-tracking method based solely on cornea images captured from a single camera.

The remainder of this paper is structured as follows. First, we briefly introduce a geometric model derived from the main characteristics of the human eye. Second, we describe how to build a 3D eye model from an image of the eye and estimate both its location and orientation relative to the camera. Third, we propose a method to continuously track the gaze direction using the previously built model. Fourth, we present the experimental results obtained using a high-sensitivity camera in low-light condition as well as prototype glasses. Fifth and finally, we conclude by suggesting further areas of work to investigate.

Eye modelization

This section describes the main characteristics of the human eye and how to derive a geometric model from them.

Human eye

Figure 1 shows a cross-section of the human eye. When observing an eye from the outside, the most distinctive parts are the colored iris, the pupil at its center and the white sclera that surrounds it, as described in (4). The outer layer of the front of the eye is the cornea, which is more difficult to observe. It covers the iris and fades into the sclera at the limbus. The cornea and lens focus images onto the retina, or more precisely, onto the fovea which is the most sensitive part of the eye. Important properties of the cornea are its transparency and its specular reflection characteristics due to the film of tears that coats its surface. This mirror-like characteristic will be particularly relevant for extracting information about the scene and the point of regard (POR).

Eye geometric model

The human eye can be subdivided into two overlapping spheres of different sizes: a smaller sphere that includes the cornea, the iris, the pupil and the lens, and a bigger sclera sphere that includes the sclera, the vitreous humor and the retina with its fovea. The two spheres intersect at the limbus which defines a circle. This model is described in Figure 1:

- Points $\boldsymbol{L}$, $\boldsymbol{C}$ and $\boldsymbol{S}$ are respectively the limbus, cornea and sclera centers. A priori unknown.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 19–24

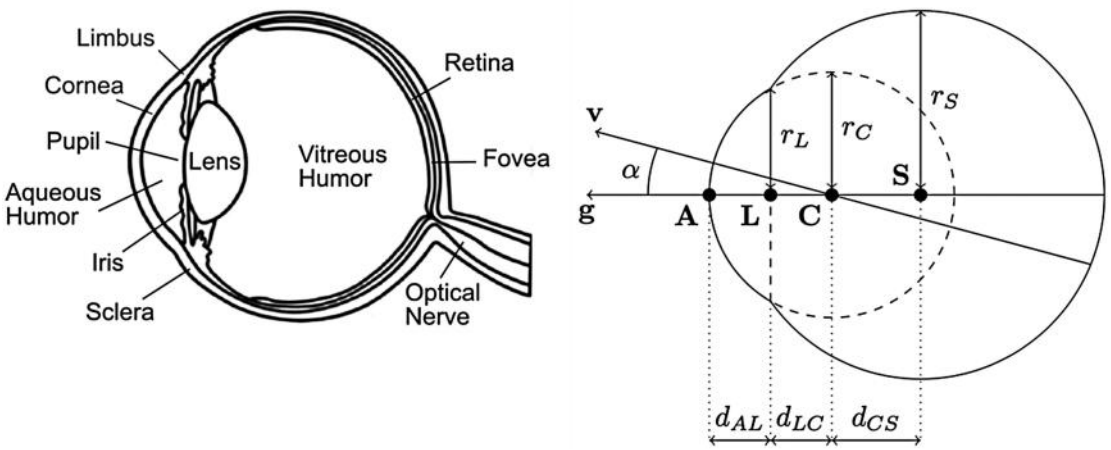


Figure 1 *Cross-section (4) and geometric model of the human eye*

• The vector $\boldsymbol{g}$ is the optical axis of the eye, crossing all the aforementioned centers. It intersects the cornea sphere at the cornea apex designated by $\boldsymbol{A}$.
• The vector $\boldsymbol{v}$ is the visual axis that goes from the fovea to the actual POR. The visual axis corresponds to the gaze direction and its estimation is the purpose of any gaze-tracking system.
• Distances r_L, r_C and r_S are respectively the limbus, cornea and sclera radii. Anatomical parameters.
• Distances d_{AL}, d_{LC} and d_{CS} separate the different components of the model. All are known anatomical parameters.

The optical axis is easy to estimate from the geometric properties of the eye but the visual axis is not. However, even though the visual axis, not the optical axis, corresponds to the direction of the POR, the optical axis can be used as a first approximation of the visual axis. The angle formed by the two axes is denoted by α and assumed to be constant.

This geometric model will be applied throughout this paper to estimate the pose of the eye from an image. It will also be used to describe interactions between the incident light and the cornea surface. By nature, such a model can only approximate the reality: the actual shape of the eye is more complex than the one described by the model and its anatomical parameters vary between individuals. However, we will assume that the variation of parameters among individuals is sufficiently small.

Eye pose estimation

Now that the geometric model is defined, we build in this section a 3D model of the eye and estimate both its location and orientation relative to the camera.

Ellipse fitting

We assume weak perspective projection since the depth of the tilted limbus is much smaller than the distance between the eye and the camera, as initially proposed in (4). Thus, the almost circular limbus projects to an ellipse described by five parameters: the center coordinates (c_u, c_v), the radii r_{max} and r_{min}, and the tilt ϕ as shown in Figure 2. Their values are estimated by fitting an ellipse on a set of limbus points, automatically detected or manually inputted, using least squares.

Pose estimation

Now that the limbus has been fully described on the image plane, we can reconstruct a 3D model of the eye and estimate its pose in the world coordinate frame, i.e. estimate the coordinates of the limbus center $\boldsymbol{L}$ and the direction of the optical axis $\boldsymbol{g}$. The following geometric construction was originally proposed by (4). The origin

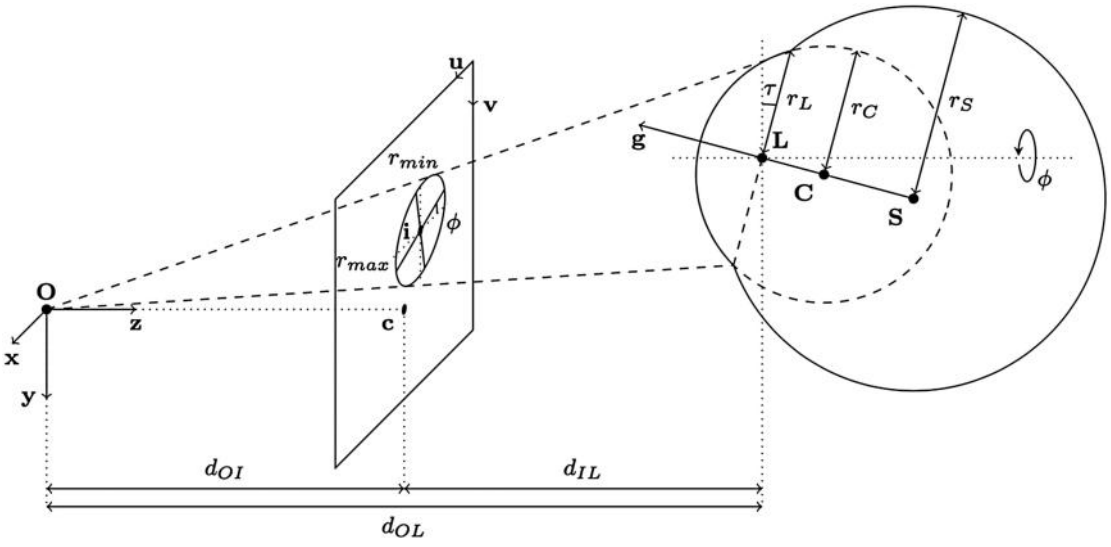


Figure 2 *3D eye model construction*

The Best of IET and IBC 2016–2017, 2016, Vol. 8,pp. 19–24

$\boldsymbol{O} = (0,\ 0,\ 0)^{\mathrm{T}}$ is at the center of the camera lens as shown in Figure 2.

When the camera focus is assumed to be at infinity[1], d_{OL} can be expressed as:

$$d_{OL} = f\frac{r_L}{r_{max}}$$

where r_L, r_{max} and the focal length of the camera f are known.

If the limbus center is defined as $\boldsymbol{L} = (L_x,\ L_y,\ L_z)^{\mathrm{T}}$, we have by similarity:

$$\frac{L_x}{(i_u - c_u)s_x} = \frac{d_{OL}}{f},$$

$$\frac{L_y}{(i_v - c_v)s_y} = \frac{d_{OL}}{f},$$

where s_x and s_y are the pixel-to-world-unit scaling coefficients, obtained from camera calibration, respectively along the x and y directions.

By combining these equations, we have:

$$\boldsymbol{L} = \left(\frac{d_{OL}(i_u - c_u)s_x}{f},\ \frac{d_{OL}(i_v - c_v)s_y}{f},\ d_{OL}\right)^{\mathrm{T}}.$$

The tilt τ of the limbus plane with respect to the image plane is estimated from the shape of the ellipse up to a sign ambiguity:

$$\tau = \pm\arccos\left(\frac{r_{min}}{r_{max}}\right).$$

Indeed, two different limbus poses are possible from the projection alone: one looking in the direction of positive values of y, and another looking in the direction of negative values of y. In the case of a head-mounted camera, the ambiguity can be easily solved by knowing the relative pose of the camera to the eye, which is usually fixed and sufficiently tilted to avoid any ambiguity[2].

The optical axis $\boldsymbol{g}$ is then given by:

$$\boldsymbol{g} = (\sin\tau\sin\phi,\ -\sin\tau\cos\phi,\ -\cos\tau)^{\mathrm{T}},$$

where ϕ is already known as the rotation angle of the ellipse fitted on the limbus in the image plane.

Finally, the cornea center $\boldsymbol{C}$ and the sclera center $\boldsymbol{S}$ are given by:

$$\boldsymbol{C} = \boldsymbol{L} - d_{LC}\boldsymbol{g},$$

$$\boldsymbol{S} = \boldsymbol{L} - (d_{LC} + d_{LC})\boldsymbol{g},$$

and the limbus is computed as the intersection between the cornea and sclera spheres.

[1]Which means $f = d_{OI}$. At close range, this sometimes cannot be assumed depending on the sizes of the sensor and the lens (6). To solve this problem, we can use a thin lens model:

$$\frac{1}{f} = \frac{1}{d_{OI}} + \frac{1}{d_{OL}},$$

where $f \neq d_{OI}$ and d_{OL} is required to compute d_{OI}. In the case of a head-mounted device, d_{OL} can be assumed to be known and constant.

[2]If not, we can also solve this ambiguity by attaching two LEDs to the camera (12). A line between the camera origin $\boldsymbol{O}$ and the cornea center $\boldsymbol{C}$ intersects at a point where the camera origin is reflected on the cornea. Therefore, we can obtain the projected cornea center location by finding the mean point of the two LEDs reflected in the image and use this information to resolve the sign ambiguity of τ.

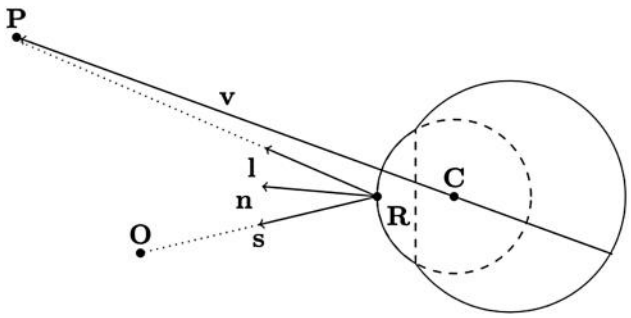


Figure 3 *Visual axis calibration*

Visual axis calibration

To evaluate the direction of the visual axis $\boldsymbol{v}$, an additional calibration step is required. Figure 3 describes the relationship between the visual axis and the incident light coming from the POR, where $\boldsymbol{P}$ is the POR and $\boldsymbol{R}$ the reflected POR, i.e. the POR in the scene reflected on the cornea image. $\boldsymbol{n} = x\boldsymbol{s} + y\boldsymbol{l}$ is the normal at the reflected POR with $\boldsymbol{l}$ and $\boldsymbol{s}$ respectively the directions to the POR and to the camera optical center. Normal parameters x and y are unknown.

In the current implementation, the user is asked to manually register the reflected POR on the image. Assuming that the distance from the reflected POR to the POR is known during the calibration process, the direction of the visual axis can be computed using a specular model of a sphere (7).

The first step is to compute the normal $\boldsymbol{n}$ in order to find $\boldsymbol{R}$. This consists of solving the following biquadratic equation:

$$4cdy^4 - 4dy^3 + (a + 2b + c - 4ac)y^2 + 2(a - b)y + a - 1 = 0,$$

where $a = \boldsymbol{s}\cdot\boldsymbol{s}$, $b = \boldsymbol{s}\cdot\boldsymbol{l}$, $c = \boldsymbol{l}\cdot\boldsymbol{l}$, $d = \boldsymbol{s} \times \boldsymbol{l}^2$ are the coefficients. When $x = (2y^2 + y + 1)\ /\ (2by + 1)$ is defined, the normal $\boldsymbol{n}$ is computed from the solution in $x > 0$ and $y > 0$. The reflected POR $\boldsymbol{R}$ and the visual axis $\boldsymbol{v}$ are then computed by straightforward vector geometry.

Gaze tracking

In order to track the optical axis direction from the current image of the eye, we first attach a pitch-yaw-roll reference frame to the sclera center $\boldsymbol{S}$ of the 3D model built previously. The yaw axis is aligned with the eye corners for convenience. The pitch and yaw angles are respectively denoted by θ and ψ. By rotating the model around the pitch and yaw axis, we simulate several limbus projections for different eye poses, as shown in Figure 4. Note that we do not consider the roll angle assuming the human eye is not capable of such a rotation.

The limbus of the rotated model is then projected into a binary image to serve as a mask. The projected area is set as binary 1s. The projection that matches the current limbus pose is then detected by summing the logical products of the inverted binary image of the current frame and the mask for each pitch and yaw values. The maximum value among the summed logical products corresponds to the current pose of the eye.

This method was initially proposed in (8) and (5). The current implementation moves toward a maximum using a Greedy algorithm. Note that the reflections on the iris area can result in white spots on the binary image that may introduce errors when computing the sum of the logical products.

Experimental results

This section presents the experimental results obtained using a high-sensitivity camera in low-light condition as well as prototype

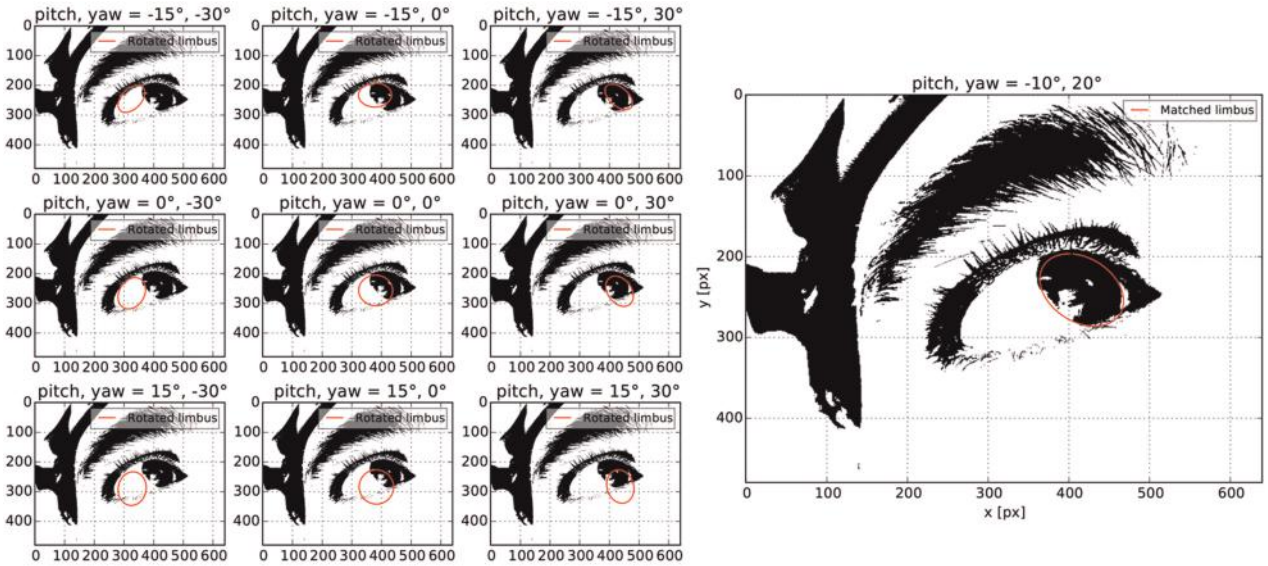


Figure 4 *Projection and matching of the rotated limbus into the image plane*

glasses. Our solution is implemented using OpenCV 3.1 C++ functions wrapped by a Python frontend. OpenCV CUDA modules are called whenever possible to benefit from GPU acceleration.

Object recognition

In order to detect the focused object, we combine the gaze direction obtained from the tracking with object recognition from cornea reflections. We propose a method based on matching 2D features between the corneal images and a reference image:

- First, we detect features and extract descriptors using Speeded-Up Robust Features (SURF).
- Then, we match the descriptor vectors using Fast Library for Approximate Nearest Neighbors (FLANN).
- Finally, we remove outliers using Random Sample Consensus (RANSAC).

Figure 5 shows the result. The detection is not perfect and suffers from noise due to iris contamination and distortions of the reflection. Further strategies must be applied to isolate the object from incorrect matches. However, our current implementation using this method runs in real-time.

High dynamic range

Cornea images are highly sensitive to light conditions and relying exclusively on a single RGB camera at night is challenging. To

Figure 5 *Object recognition using 2D feature matching*

The Best of IET and IBC 2016–2017, 2016, Vol. 8,pp. 19–24

Figure 6 *HDR processing result (right) in low-light condition*

address this issue, we conducted an experiment using a ViewPLUS Xviii high-sensitivity camera (9) and High Dynamic Range (HDR) processing at night time. The Xviii camera captures eleven 8-bit RGB images at increasing levels of sensitivity over an 18-bit dynamic range. We combine these images using a HDR algorithm, such as Exposure Fusion, to reveal the features reflected on the cornea. Figure 6 shows the result of a user watching a computer screen (displaying the reference object of Figure 5) in the darkness. By using a high-sensitivity camera combined with HDR techniques, we are able to apply our proposed gaze-tracking method in low-light condition.

Prototype glasses

To assess the precision of our tracking method, we prototyped a head-mounted device using a pair of JINS MEME glasses (10) as a base. We mounted on top of them a 3D-printed frame to fix the cameras used for corneal imaging, as shown in Figure 7. We use two e-con Systems See3CAM_80 RGB cameras (11). Eye images from both eyes can be captured at a near-4K resolution of up to 3264×2448 pixels at 11 frames per second. The two video streams are passed to a computer through two USB 3.0 cables via an USB Video Class (UVC) 1.1 standard interface.

During the experiment, a user is asked to sit in front of a computer screen and look at several targets displayed at each corner while staying still, as shown in Figure 7. The camera position relative to the screen is measured and assumed to be constant. The visual axis direction is computed from the measurements and compared with the one given by the implementation of the proposed method.

Results obtained with a screen placed at 500 mm indicate a mean error of 2.5° and a maximum error of 5°. Most commercial infrared-based solutions advertise higher precision, usually below 1°. However, they exclusively focus on gaze direction estimation

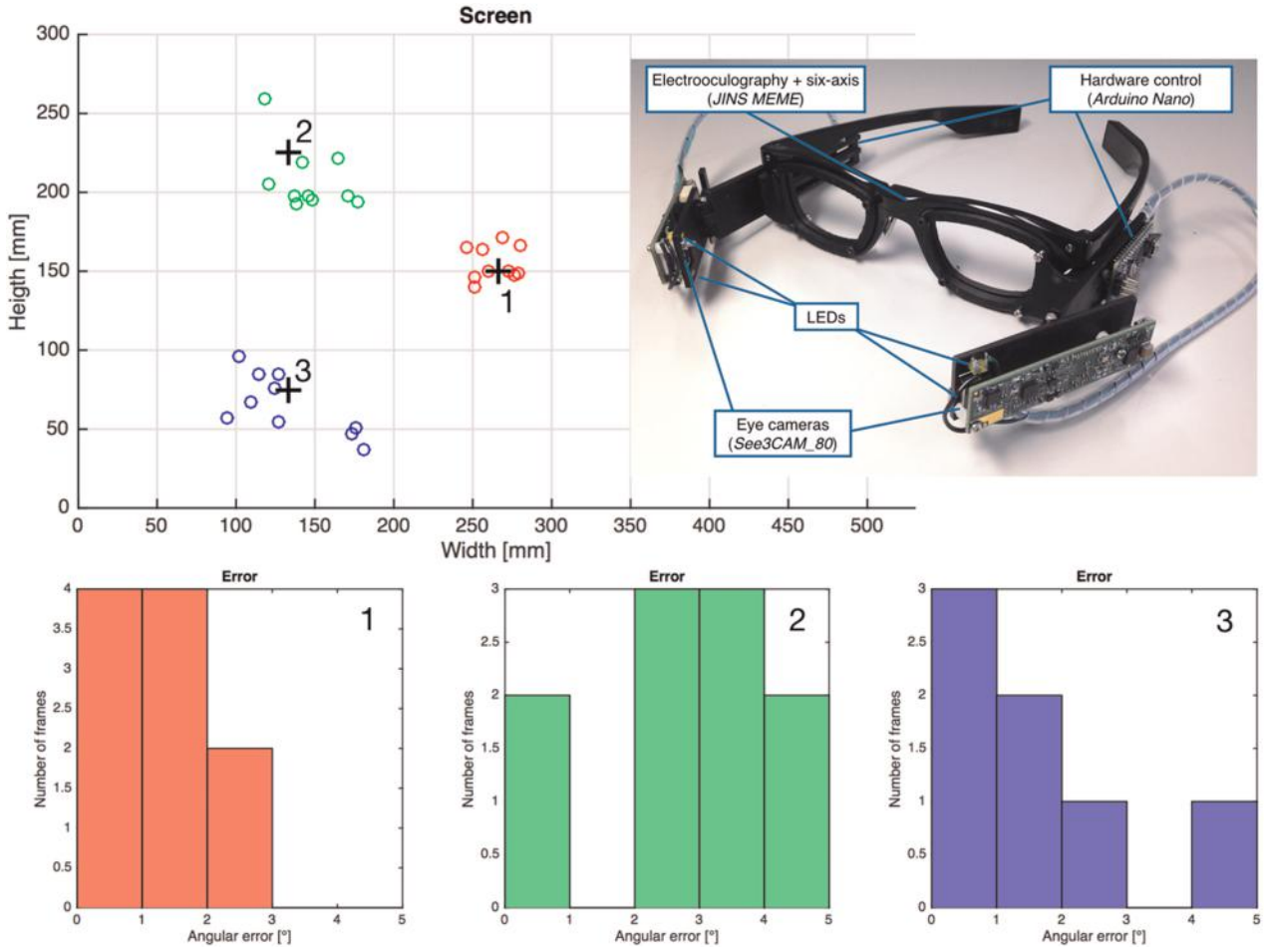


Figure 7 *Experimental results using prototype glasses*

and require additional sensors to extract scene information. Moreover, they do not work well outdoors because of the sun. Hence, targeting an angular error of less than 3° seems reasonable when using natural images prone to more noise than infrared ones, especially considering that extracting scene information at the same time as gaze direction could make up for the lack of precision, depending on the application.

Conclusion

This work explored the feasibility of wearable eye-tracking systems solely relying on a single camera to evaluate the direction to the POR and recognize objects reflected on the cornea surface. However, cornea images are highly sensitive to light condition and many difficulties remain, such as the calibration of the visual axis and the individual variations of the 3D model parameters. Future work will include:

- Extracting the anatomical parameters of the user to improve accuracy.
- Removing the extra step of visual axis calibration.
- Evaluating not only the direction of the gaze, but also the depth to the POR. This could be achieved with stereo reconstruction from left and right cornea images.
- Unwrapping cornea images from the reflected POR to remove distortions before applying object recognition algorithms.

Hence, eye tracking based on corneal imaging requires further investigations to address these challenges in order to become convenient enough for daily life purposes.

Acknowledgments

This work was conducted under the MEXT scholarship program of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1 Geuss, M.: 'Why eye tracking could make VR displays like the Oculus Rift consumer-ready', *Ars Technica*, 2014
2 Nitschke, C., Nakazawa, A., Takemura, H.: 'Corneal imaging revisited: an overview of corneal reflection analysis and applications', *IPSJ Trans. Comput. Vis. Appl.*, 2013, **5**, pp. 1–18
3 Nishino, K., Nayar, S.K.: 'Corneal imaging system: environment from eyes', *Int. J. Comput. Vis.*, 2006, **70**, (1), pp. 23–40
4 Nitschke, C., Nakazawa, A., Takemura, H.: 'Display-camera calibration using eye reflections and geometry constraints', *Comput. Vis. Image Understand.*, 2011, **115**, (6), pp. 835–853
5 Takemura, K., Yamakawa, T., Takamatsu, J., Ogasawara, T.: 'Estimation of a focused object using a corneal surface image for eye-based interaction', *J. Eye Move. Res.*, 2014, **7**, (3), pp. 1–9
6 El Hafi, L., Takemura, K., Takamatsu, J., Ogasawara, T.: 'Model-based approach for gaze estimation from corneal imaging using a single camera'. Proc. IEEE/SICE Int. Symp. on System Integration (SII), 2015, pp. 88–93
7 Eberly, D.: 'Computing a point of reflection on a sphere', 2008, pp. 1–4
8 Takemura, K., Kimura, S., Suda, S.: 'Estimating point-of-regard using corneal surface image'. Proc. Symp. on Eye Tracking Research and Applications (ETRA), 2014, pp. 251–254
9 ViewPLUS: Xviii, 2015, http://www.viewplus.co.jp/product/camera/xviii.html
10 JIN: JINS MEME, 2015, https://jins-meme.com/
11 e-con Systems: See3CAM_80, 2013, https://www.e-consystems.com/8MP-USB3-Autofocus-Camera.asp
12 El Hafi, L., Uriguen Eljuri, P., Ding, M., Takamatsu, J., Ogasawara, T.: 'Wearable device for camera-based eye tracking: model approach using cornea images'. Proc. JSME Robotics and Mechatronics Conf. (ROBOMECH), 2016

The Best of IET and IBC 2016–2017, 2016, Vol. 8,pp. 19–24

Creating object-based experiences in the real world

Michael Evans, Tristan Ferne, Zillah Watson, Frank Melchior, Matthew Brooks, Phil Stenton, Ian Forrester

BBC Research and Development, UK

Abstract: The move towards end-to-end IP between media producers and audiences will make new broadcasting systems vastly more agnostic to data formats and to diverse sets of consumption and production devices. In this world, object-based media becomes increasingly important; delivering efficiencies in the production chain, enabling the creation of new experiences that will continue to engage the audience and giving us the ability to adapt our media to new platforms, services and devices. This paper describes a series of practical case studies of our work in object-based user experiences since 2014. These projects encompass speech audio, on-line news and enhanced drama. In each case, we are working with production teams to develop systems, tools and algorithms for an object-based world: these technologies and techniques enable its creation (often using traditional linear media assets) and post-production; transforming user experience for audiences and production.

Introduction

In 2014 BBC R&D presented an IBC paper on object-based broadcasting [1], the representation of media content by a set of individual assets together with metadata describing their relationships and associations, and the abilities to bring these back together again to make new content experiences. This work has continued to progress since those very early prototypes and proofs-of-concept. We have now created a range of object-based experiences, together with experimental tools to enable the sustainable creation of such content. Collectively, these systems have formed a valuable catalyst for building our knowledge and understanding of how producers of creative content can design and deliver these experiences. We find them to be useful, practical case studies that should enable broadcasting organisations to thrive among the new broadcasting systems evolving from end-to-end IP and ubiquitous computing.

In each of the scenarios described here, we have worked with production teams to develop systems, tools and algorithms for an object-based world: these technologies and techniques enable its creation (often using traditional linear media assets) and post-production, transforming user experiences for audiences and enhancing the craft of production professionals. In [1] we emphasised the continued importance of skilled craft in the curation of audio and video objects, as well as the data objects that describe them and their relationships and roles in the audience experience. This included the construction of a layered curatorial model, relating richer description of content relationships to more responsive experiences. In this paper we will see this distilled into the craft and the opportunities in curating the semantics of objects, exploiting descriptive relationships between experiences and the elements comprising them. Specifically, this paper describes the following projects:

- Discourse – a text-based semantic editing system for audio production
- Atomising News – structured storylining of content to support dynamic presentation
- Squeezebox – a tool for adding prioritisation semantics to segmented linear content, to allow simple control of content duration
- StoryExplorer / StoryArc – presenting an interactive experience based on the semantics of a drama and assisting writers in their craft.
- Visual Perceptive Media – a pilot of a richly annotated set of video assets, which can be assembled into a short drama based on each viewer's current context.

Discourse: semantic audio editing

Speech radio listenership remains high and podcasting continues to grow in popularity. Although much speech content is still broadcast live, a large proportion is pre-recorded and the experience constructed using audio editing software. Commonly, such tools represent sound using simple waveforms, allowing users to visually search and scan audio content but displaying very limited information. This approach does not scale well [2]. Efficient navigation and editing of speech is crucial to the radio production process. However, unlike text, speech audio must be navigated sequentially and does not naturally support visual search techniques [3]. Furthermore, the authoring of object-based experiences may also require the annotation of the speech audio with semantic mark-up describing various useful attributes; functionality not generally offered by waveform editors.

Semantic analysis techniques can be used to extract higher-level information from the audio, such as: whether the content is speech or music [4], where different people are speaking [5] or a transcript of what they are saying. Presenting *this* information to the user could allow them to navigate and edit audio content much more efficiently. They can also be used to create new experiences like responsive radio or variable-length programmes.

Over the last year *Discourse* has been developed; a semantic audio editing system that uses a text-based interface to enable users to navigate and edit speech using an automatically generated transcript. Development included a qualitative study of current radio production and evaluation of semantic editing. We found that current practice involves time-consuming note-taking and logging, before editing the audio based on these notes. The semantic editing system allows producers to complete this process up to twice as fast in some cases. However, the semantic system was not as efficient for short recordings. Participants commented that *Discourse* allowed them to navigate and edit the audio much faster and that the accuracy of the transcript was good enough for their purposes. Both of these results support previous findings [6, 7].

It is also important to acknowledge that editing decisions are based not only on what is said, but how it is said. This emphasises the importance of a multi-modal interface, which combines the efficiency of text-based working with being able to quickly listen to the audio.

The system—which is now in the process of becoming a BBC internal product—was designed to offer basic functionality for creating a rough edit, and to easily fit into the existing production

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 25–29

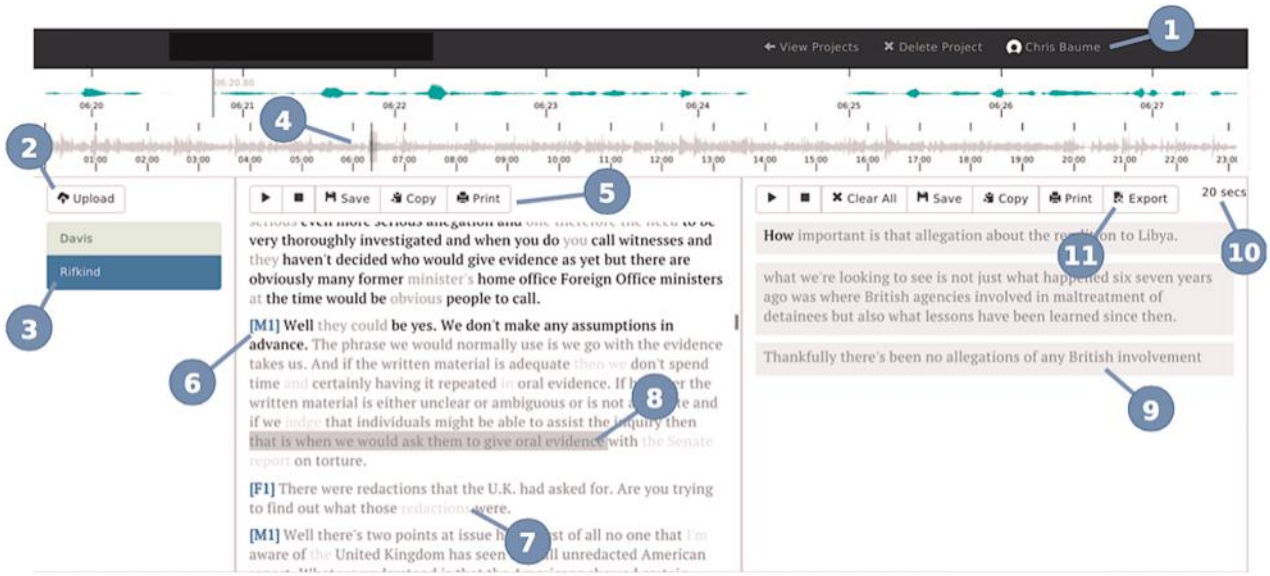


Figure 1 *Discourse UI: (1) user accounts and projects, (2) upload audio recordings, (3) upload list, (4) waveform display of currently selected recording, (5) toolbar with playback, save, copy and print functions, (6) transcript of selected recording with speaker labelling and word editing, (7) confidence shading, (8) transcript selection with drag-and-drop editing, (9) listing and re-ordering of edits, (10) duration of edit, (11) export edit.*

workflow. A screenshot of the interface and numbered list of the main features are shown in Figure 1.

The effectiveness of incorporating semantics into craft processes highlights audio tools and experiences as being a core part of researching object-based broadcasting.

ATOMISING NEWS

In a similar manner to speech audio, some forms of news and journalism (regardless of platform) also lend themselves well to semantic representation and atomisation as media objects. The Storyline Ontology (http://www.bbc.co.uk/ontologies/storyline) developed by the BBC and others, defines storylines as being made up of events where each event can be linked to people, organisations, places and even other storylines. Although most news is complicated and subjective (Where do stories start or end? Where do they intersect? What really happened?) these structures can be used to create new media experiences.

By splitting news stories into these "atoms" of events and actors and fitting the atoms to a data model we create re-usable bits of stories that we know when and how to use, and then can put together again in different ways. These structures give us the ability to adapt our stories to different screens, platforms, contexts and experiences. News stories can be optimised for users of smartwatches, conversational interfaces, or futuristic agents using AI. Stories can also be personalised to users, which is particularly important for news, media and broadcast organisations. Audiences use an increasing range of platforms and devices and, rather than create more and more content for each of these new platforms piecemeal as they appear, if we structure our stories into small, reusable pieces then we can efficiently re-use them again and again.

Our initial work based on these structures in news is aiming to develop a news format suitable for a younger audience and for mobile users. Through several iterations of prototyping and testing [8] we discovered that people liked having quick, skimmable summaries of stories, but they also wanted the option to be able to go deeper and get more information on the aspects that interest them.

Our prototype presents every story structured as a storyline with key events, people and places. The initial view of the story gives you a summary on one page, so you can just skim the key events. But you can expand any of the events and dig deeper — into longer-form writing, correspondent reports, social media or video. The prototype also includes pop-up definitions for the key people, organisations and places mentioned in the story. Further pilots planned for 2016 will build understanding about how journalists can write for this format.

Squeezebox

Squeezebox is a prototype production tool that explores automated video re-editing using semantic mark-up. News and other factual content is often needed at different durations for use in different programmes. Edits will be made manually, and can be time-consuming and laborious. *Squeezebox* aims to assist in the rapid re-editing of such content, allowing new durations to be instantly produced.

Currently, we target the production of captioned news story montages. Using the tool, production users add simple metadata to a collection of news stories and caption them. News story montages can then be instantly produced by selecting a target duration. Footage uploaded to *Squeezebox* is automatically analysed and segmented into individual shots. Then, rather than manually editing the footage, the user marks-up the most relevant and important portions of each shot, indicating that the rest is a candidate for being cut. She also marks up the priority of each shot, determining: how the footage behaves as the duration is reduced, which shots will be dropped first, and which ones will be preserved. Using this metadata, *Squeezebox* enables users to adjust the duration of the story using a simple slider control. The purpose-built algorithm establishes new in-and-out edit decisions per shot and in some cases drops shots entirely.

Captions are also prioritised and an appropriate number of captions are chosen to fit a story's duration, ensuring that a maximum reading rate is not exceeded. A music bed can also be selected. As we automatically transcribe any dialog found in news story videos, it's possible to use text selection to highlight phrases and instantly create "*upsounds*" – which, if chosen, automatically duck the music bed volume, restoring it when the phrase ends. User-specified idents automatically top-and-tail the montage.

Production user research

Empirical research found that users liked the *Squeezebox* concept, with most users seeing future potential and finding the tool easy to learn and use. Satisfaction levels with the montages produced by *Squeezebox* varied: one user with experience of editing in

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 25–29

time-critical newsroom environments appreciated the speed at which they could re-edit content, and was happy to cede precise control to the tool. Another user who didn't normally edit news wanted tighter control than *Squeezebox* allows.

In terms of application, users indicated that it might be suitable for use by teams who don't necessarily have the knowledge or budget for high-end professional tools. It could effectively up-skill non-editors, enabling them to create finished packages without needing to involve other craftspeople. It could also allow resource-constrained teams to reduce the overheads of producing additional social media content on top of their existing output.

Story explorer

Drama on TV frequently has a complex narrative and many of us have had the experience of watching part of a long-running series whilst thinking "*Who was that?*" or "*What just happened?*", or forgetting a crucial event an earlier series. In effect, viewers periodically crave supplementary narrative semantics. Using object-based media we developed something to help solve this problem - for people to recap, catch-up and explore the stories from TV and radio drama.

We piloted the concept with a daily radio drama, Home Front, in Summer 2015 on the BBC's Taster platform for audience experiments. The prototype website that we built, the Home Front Story Explorer (Figure 2, see also [9]), lets listeners choose the storyline with which they want to catch up, and they can then listen to that entire storyline, or highlights, or skim through it and read the parts that interest them. They can see which characters are in each scene and find out more about them. This is not intended to be a replacement for listening to the programme, but to be useful for people who want to quickly catch up or remind themselves, or who want an introduction to a drama that's new.

To construct the data and media objects for this, we modelled the stories and designed the application around the stories rather than series and episodes. We used three layers to model stories:

1) The story world, consisting of the people (protagonists), places, and relationships;
2) The events that happened in this story;
3) The narrative: how the story is told.

Based on our research with users, the primary building blocks we wanted for the *Story Explorer* were storylines, key events and people. We took the first two series of the drama and defined the key storylines, the events or scenes in each storyline, the characters and the places involved. Then for each scene and character we snipped-out the audio from the appropriate episode and wrote a textual summary, creating a self-contained, self-described "atom" of media for every scene and character.

The pilot was very successful with Home Front's radio audience. As expected, some were new to the drama and used the *Explorer* to get into the story. Some just wanted a recap on bits that they had missed or forgotten. Some even used it as a single-purpose radio — listening to single storylines for hours!

Our work now continues on refining and simplifying the user experience, making re-usable design patterns that can be used across radio and TV dramas [R&D blog post coming today or tomorrow] and exploring other platforms and interfaces - starting with mobile and moving on to TVs.

Storyarc

One of the challenges for the BBC in developing a new experience like the *Story Explorer* is how to scale it across the many dramas that we produce every year. Therefore, we have been working on tools

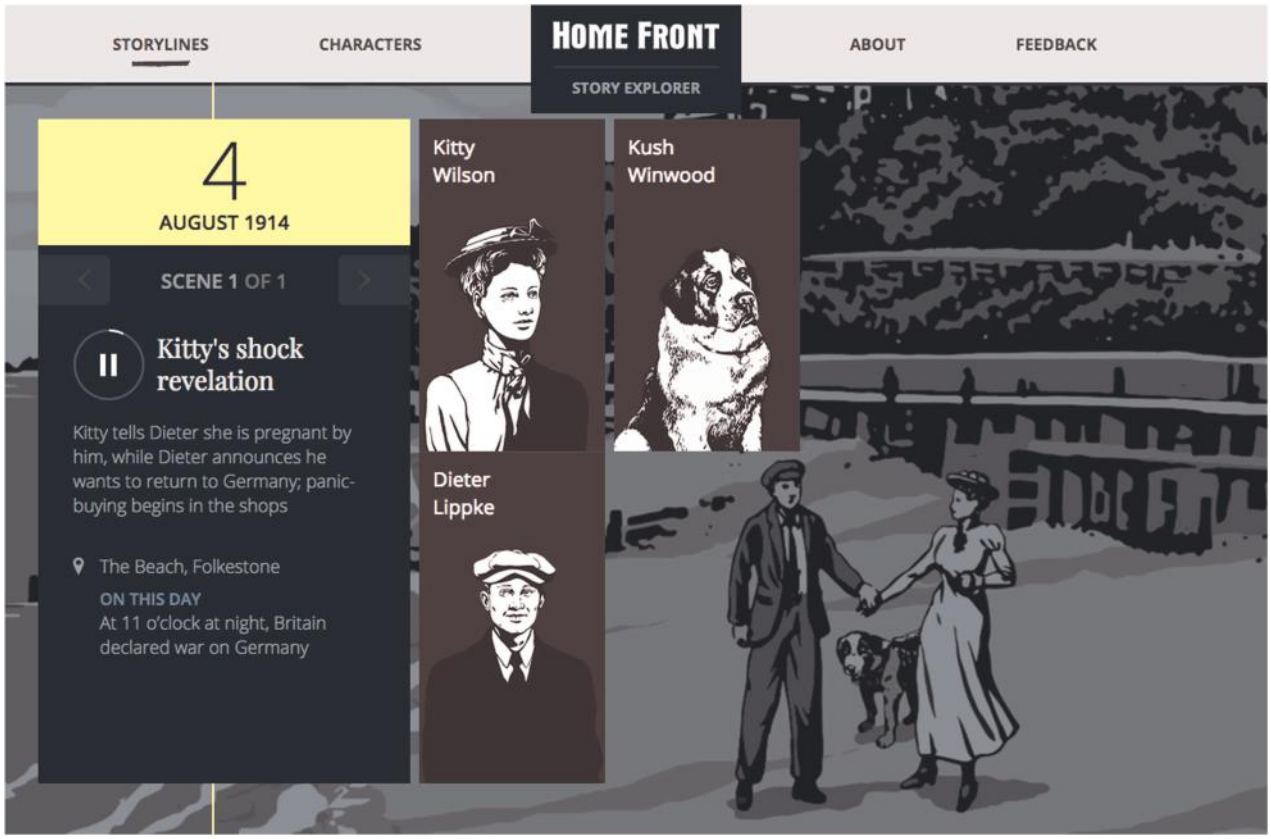


Figure 2 *The Home Front Story Explorer showing a scene and the characters involved*

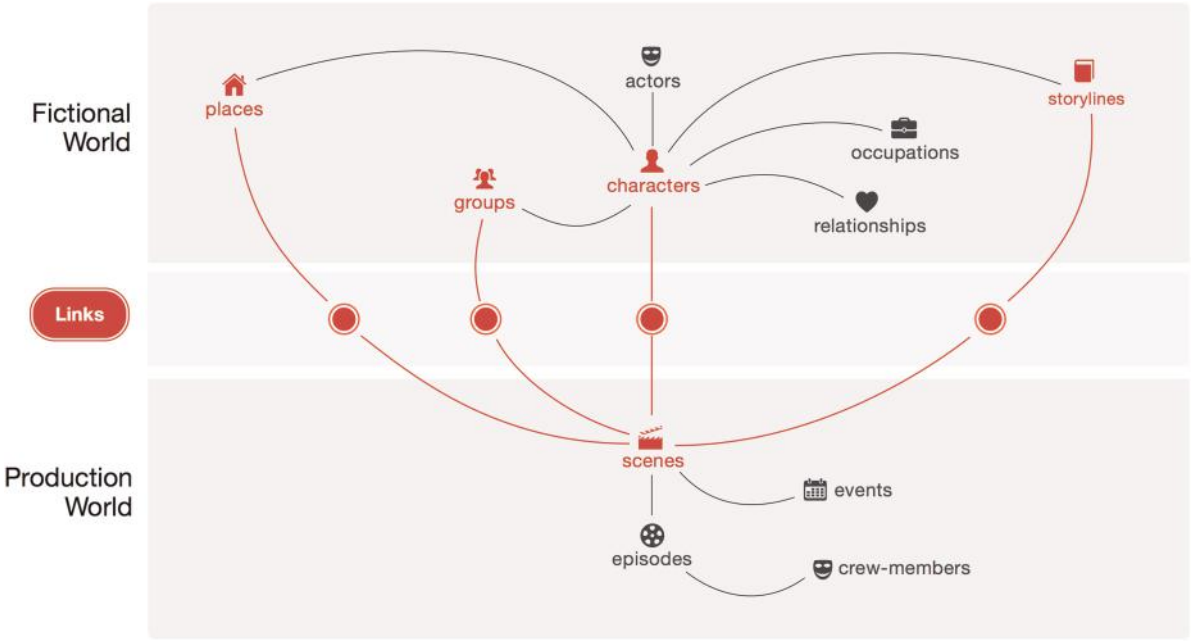


Figure 3 *The Storyarc model showing how the data links production and fictional worlds*

that might be used to efficiently generate the structured story data and object-based media as part of existing production processes.

The Archers, 65 years old this year, is the world's longest running soap opera. It is about a farming community and is broadcast every day on UK radio. The production has had a fairly rudimentary continuity database and an archivist for a number of years: recorded are such data as: what happened in 1967; who married who, who's got a piano in their front room and even what the cows are called! This is used by the writers to be consistent and to give them inspiration for new storylines. We have designed a new web-based structured continuity database called *Storyarc* that replaced this old system, based upon our object-based story model. *Storyarc* is illustrated in Figure 3.

Storyarc is now in daily use by the researchers, writers and producers in *The Archers* team. It is a good example of a tool that has been integrated into existing production workflow and become useful to the team in their day-to-day work. The structured story data and object-based media produced as a by-product, can then be used to create new experiences. We deliberately designed the data model to serve both production and audience-facing needs and *Storyarc* capitalises on this. We are now looking at ways to make the authoring even easier by automatically processing existing media, including machine-parsing programme scripts to create the story data semi-automatically.

Visual perceptive media

In this project we are researching the value that object-based content can bring, by composing each user's response to the context at the moment of delivery. If we can create content that can be composed just before or during delivery we might be able to use the semantics describing each viewer's requirements and context in order to present the most useful, enjoyable or salient experience. For example, educational content that could respond to each viewer's skill level and the challenges they encounter. Such a programme could be configured (composed) to present the right level of information and insert extra emphasis on the things *you* need to learn or see demonstrated at the time *you* need them. Perhaps the intelligibility of a particular character's dialogue requires the volume for just that character to be increased or subtitles to be included.. People who have poor sight might want the contrast of a scene to be increased to see the detail better. The main challenges are fourfold:

- How do you tag the content in a way that allows multiple re-compositions of the objects?
- What production craft is necessary to capture and produce content that might tell the same story in different ways (shorter, longer, more or less romantic…)?
- Which contextual attributes are important in order to drive composition?
- What are the requirements of the delivery engine, to support flexible composition?

As a pilot we created media assets for a simple responsive storyline, filming all the shots required to enable the flexibility composition. Audio and video were captured and stored as independent objects and a number of 'look-up tables' were constructed for colour grading. From these objects, 32 variations of a five-minute short film can be rendered algorithmically, based on the preferences and context of the audience. (Variation includes a male or female perspective, different colour grading and changes in comedic/dramatic tone.) In essence the film does not exist until it is shown; assembled just-in-time by selecting and timing individual clips, grading and soundtrack objects, from a repository. Our ongoing research questions the challenges, opportunities and potential value of these experiences and examines how to enable a "community of practice" in storytelling and production craft in this domain.

Conclusions

We have described several case studies of recent work in the production of object-based experiences. It is clear that curating audio, video and data assets with specific meaningful semantics to describe their roles and interrelationships, has the potential to transform user experiences for both audiences and producers. There is a range of interesting challenges that need to be investigated in order to deliver this value: Which kinds of semantic data are the most valuable in facilitating the combination of media objects? What other parts of the user experience can take advantage of object-based media—in terms of new forms of content, or more efficient production and accessible production—and how should we evaluate the practical extent of this benefit? How do we further simplify the making of object-based content, and how will the required new tools fit into established

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 25–29

workflows? Critically, what is the balance between automation and human craft embodied in these craft processes?

References

1 Armstrong, M., Brooks, M., Churnside, A., Evans, M., Melchior, F., Shotton, M.: 'Object-based broadcasting – curation, responsiveness and user experience'. Proc. 2014 International Broadcasting Convention (IBC2014), 2014
2 Loviscach, J.: 'The quintessence of a waveform: focus and context for audio track displays'. 130th Audio Engineering Society Convention, 2011
3 Wolfe, J.M., Horowitz, T.S.: 'What attributes guide the deployment of visual attention and how do they do it?', *Nature Rev. Neurosci.*, 2004, **5**, pp. 495–501
4 Panagiotakis, C., Tziritas, G.: 'A speech/music discriminator based on RMS and zero-crossings', *IEEE Trans. Multim.*, 2005, **7**, pp.155–166
5 Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: 'Speaker diarization: a review of recent research', *IEEE Trans. Audio, Speech, Lang. Process.*, 2012, **20**, pp. 356–370
6 Whittaker, S., Amento, B.: 'Semantic speech editing'. Proc. ACM SIG CHI '04, 2004, pp. 527–534
7 Rubin, S., Berthouzoz, F., Mysore, G.J., Li, W., Agrawala, M.: 'Content-based tools for editing audio stories'. Proc. ACM UIST '13, 2013, pp. 113–122
8 http://www.bbc.co.uk/rd/blog/2015-09-elastic-news-on-a-mobile
9 http://homefront.ch.bbc.co.uk

Dreamspace: a platform and tools for collaborative virtual production

O. Grau[1], V. Helzle[2], E. Joris[3], T. Knop[4], B. Michoud[5], P. Slusallek[1], P. Bekaert[6], J. Starck[7]

[1]*Intel-Visual Computing Institute, Germany*
[2]*Institute of Animation at Filmakademie Baden-Württemberg, Germany*
[3]*CREW, Belgium*
[4]*Stargate, Germany*
[5]*Ncam Technologies, UK*
[6]*iMinds, Belgium*
[7]*The Foundry, UK*

Abstract: This paper describes the concepts and results implemented by the European FP7 Dreamspace project. Dreamspace is developing a new platform and tools for collaborative virtual production of visual effects in film and TV and new immersive experiences. The aim of the project is to enable creative professionals to combine live performances, video, and computer-generated imagery in real-time. In particular, the project has developed tools, allowing on-set manipulation of 3D assets, live integration of video feeds from tracked cameras, and live-compositing of either CGI content or background plates from panoramic video, captured by Omnidirectional video rigs. The CGI content is lit by automatically captured studio lighting, using a new real-time global illumination rendering system. Dreamspace is investigating the use of omnidirectional video and 3D assets in new immersive user experiences.

Introduction

The film and TV industry is seeking ways to produce audio-visual media that combines the real world, CGI, and 3D animation in ever increasing quality, but at lower cost. Using CGI in movie and TV productions has reached a degree that makes it hard to identify what parts of a production are real and which are virtual. However, the traditional two-phase approach of on-set filming and integration of visual effects, later in a post-production phase, has proven to be a major bottleneck in terms of creativity and cost-effectiveness.

The European FP7 Dreamspace project (1) has developed new techniques and workflows to provide full creative control over the virtual (computer-generated) components in production with real-time visualisation, and continuity of the data and creative decisions through to post-production. This includes: intuitive on-set manipulation of 3D assets, live camera tracking and compositing with depth for visualisation, and capture of on-set lighting and real-time global illumination rendering to harmonise real and virtual elements. All this is done with panoramic video to provide low-cost photo-real environments. The project has also explored the impact on film production and the cross-over to create new immersive experiences for live performance and installation art, using emerging head-mounted displays and head-tracked projection screens.

This paper introduces the concepts developed in Dreamspace. It combines leading research and commercial organizations in imaging, visual production, and creative experiences, having seven partners. The text provides an overview, followed by a more detailed description of data capture and processing, the collaborative virtual production environment, and an overview of some of the creative virtual productions carried out in the project.

Overview

Dreamspace has created an end-to-end pipeline for data capture, processing, and rendering, with control of virtual elements, to produce a visualisation environment for use on set in both film and TV. It can also be part of an immersive space for installation or performance art.

The technologies and prototypes focus on providing greater control, streamlining the workflow and pipeline, connecting the different phases of production, and delivering high quality content at a lower cost. These targets span virtual production and the performance arts, where the challenge is the same - enhance creativity and experimentation through real-time visualization and interaction with accessible tools and workflows.

Data capture and live visualisation

Real-time camera tracking and depth capture

In virtual production today, live-action foreground elements are visualised in real-time with a CGI background to allow the director and director of photography to make creative decisions on shot framing, lighting, and timing. Live visualisation systems require a specialist studio with a chroma-key background to separate the foreground elements and camera tracking to render the CGI background. The preview is usually restricted to a simple live-action overlay. In Dreamspace, a real-time camera tracking system with live depth capture has been developed to allow integration of live-action content inside a virtual set, with dynamic occlusions between real and virtual elements. The system works in natural scenes, potentially removing the need for a dedicated studio and green screens, to make virtual production techniques accessible to a range of productions.

Two approaches to live depth capture have been developed, using the *ncam* camera bar as the only hardware. The first prototype focuses on regular blue or green screen studios. It uses the known background to separate the live action elements and compute the corresponding depth information. This supports dynamic occlusions in existing virtual production studios. The second prototype has been designed for uncontrolled backgrounds. It iteratively computes a 3D model of the static scene geometry. As the camera is tracked, the live action elements are separated to provide depth measurements for moving elements, with the static background. This supports live depth capture in natural scenes.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 30–35

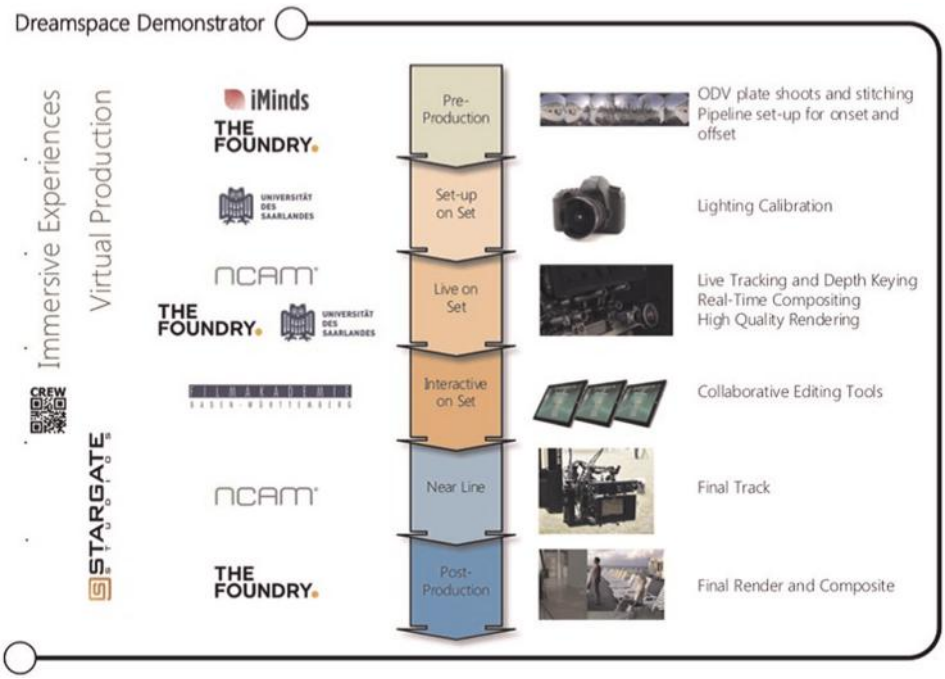


Figure 1 *Shows how the technologies fit a conventional production pipeline with capture of video back-plates using: omni-directional video camera rigs, on-set capture of light models to harmonise real and virtual lights in rendering, real-time integration of real and virtual elements through live camera tracking with depth capture, high-performance global illumination rendering and real-time compositing for live visualization, collaborative editing tools to control digital content, and, finally, data continuity through to post-production to deliver the final camera track and final composite*

The real-time system on set must be robust, reliable, accurate, easy-to-use and with minimal latency to fit within existing film pipelines. A post-processing technique has been developed to deliver a final match-move camera, ready for post-production using the data recorded from set. This removes the need for dedicated match-moving in post production.

Live on-set compositing

On-set visualisation provides a live preview to assess shot composition. Final visual effects are then typically created in post-production, using separate offline tools that allow artists complete control over the look and integration of real and virtual elements. In Dreamspace, a live visualisation system has been developed that allows the virtual set and the live composite to be prepared, using conventional post-production tools, with the same pipeline and results both on-set and in post-production.

Figure 2 *Live depth capture*

The live compositing system supports an arbitrary graph of user-defined image processing operations created in Nuke[1]. A heterogeneous scheduling system has been created to make optimal use of hardware in the on-set system. The scheduler distributes data processing to CPU and GPU devices to maximise the frame-rate. The system provides the flexibility to switch and modify the composite live on set. This is then passed to an offline pipeline, designed to deliver the final integration of real and virtual elements, based on the RGB + depth data recorded from set. An image-based matting technique has been developed to estimate automatically the opacity at the boundary between depth layers in the scene. This allows virtual elements to be integrated seamlessly, with no manual rotoscoping or green screen keying.

Real-time rendering with global illumination

Live visualisation requires real-time rendering, which typically restricts the quality that can be achieved on-set. In post-production, global illumination is used to create highly realistic virtual scenes, but at the cost of many hours to render a single frame. In Dreamspace, a framework for high-performance global illumination rendering has been developed, with a scalable distributed architecture, to achieve final quality rendering in real-time on-set.

Ray-tracing routines have been developed using AnyDSL (4), a compiler framework for domain-specific libraries (DSLs). Current state-of-the-art frameworks for rendering provide low-level routines that are, optimized and tied to a single platform, such as Embree from Intel or OptiX from NVIDIA. AnyDSL allows mapping of ray tracing routines to different hardware platforms, using code refinement and partial evaluation (4), using available

[1]http://www.thefoundry.co.uk/products/nuke/

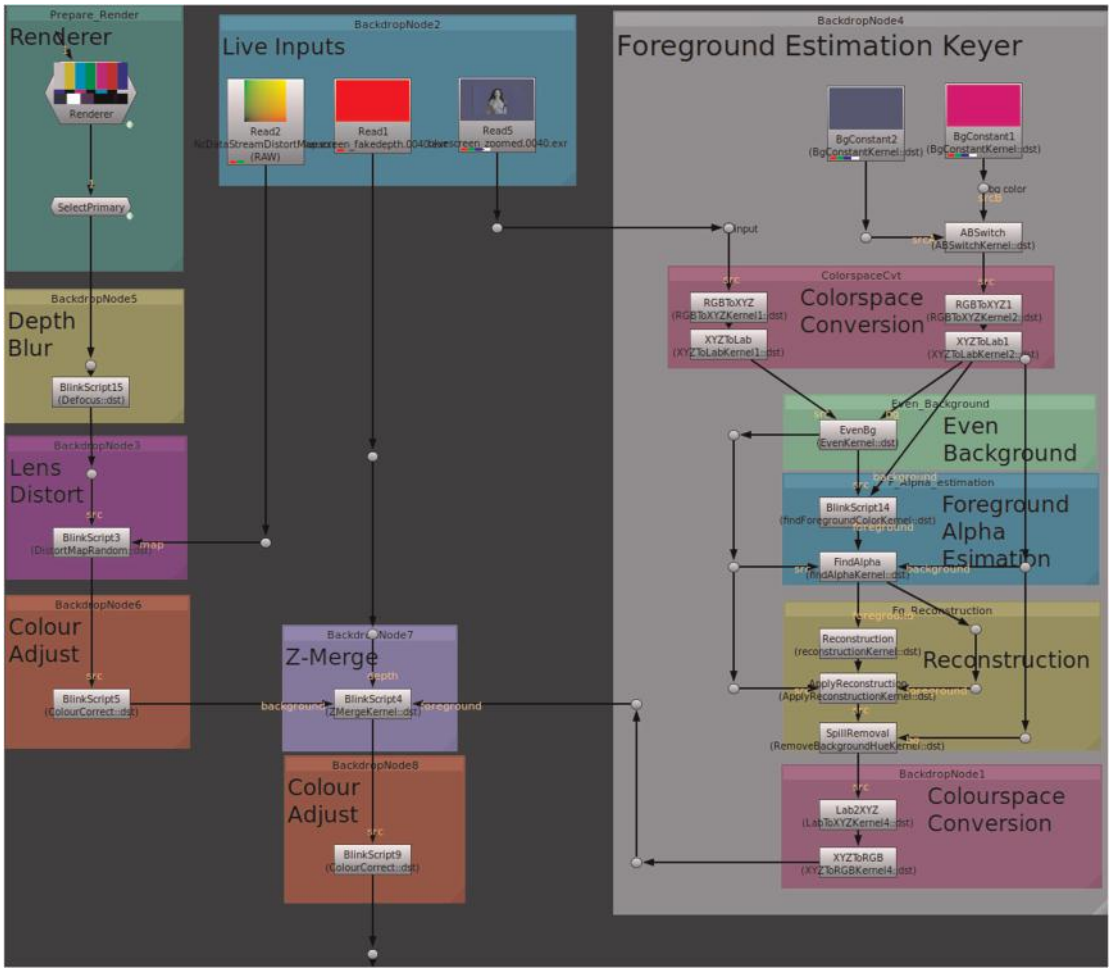


Figure 3 *Live compositing supports an arbitrary graph of user-defined operations*

hardware resources. We achieve the same performance – sometimes, even slightly faster - than the highly hand-optimized implementations from Intel and NVIDIA on a single node. Rendering is distributed over multiple nodes in a cluster with load balancing to account for varying scene complexity to achieve real-time performance. We have implemented three algorithms for light simulation on top of the low-level ray tracing routines. Figure 4 shows the scene 'San Miguel' rendered in *LiveView*, using Path Tracing. As an alternative, the renderer can also use Bidirectional Path Tracing and Vertex Connection and Merging as the rendering algorithm. It supports live scene updates coming from *LiveView*.

Figure 4 *The San Miguel scene rendered with global illumination in LiveView*

Collaborative virtual production environment

Virtual Production Editing Tools (VPET[2])

Virtual studio systems require highly technical and dedicated tools to control the on-set environment. In Dreamspace, we have developed a holistic approach for collaborative tools that allow on-set light, asset, and animation editing via an intuitive interface. The Virtual Production Editing Tool (VPET) is a tablet-based on-set editing application that works within a real-time virtual production environment. It is designed to run on mobile and head-mounted devices (HMD), allowing easy access to edit parameters of virtual objects (set, lights, animation), without dedicated training. VPET provides live editing of assets in the film pipeline, synchronized across all VPET clients and Digital Content Creation (DCC) applications through a network interface.

The client application is realized, using the Unity framework. Figure 5 (bottom) shows early versions with HMD and gesture approaches; evaluation (2) of those have led to using tablets in the current versions. Working with VPET requires no dedicated export procedure. The scene only needs to hold a medium resolution representation of the geometry for playback on mobile devices. VPET connects to the central production system, which streams the scene automatically to the clients. Authoring of content is realized through the established Katana[3] software for

[2]http://vpet.research.animationsinstitut.de/
[3]http://www.thefoundry.co.uk/products/katana/

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 30–35

Figure 5 *(upper) VPET during live production at Filmakademie and (lower) early prototype using HMD and gesture recognition sensor*

appearance (look) development and lighting. Using Katana maintains an established film production pipeline for real-time, collaborative use in virtual production. VPET is not restricted to one central production system. The scene distribution can be easily adapted to any DCC application. VPET clients can communicate to other hardware used in virtual production, such as the *ncam* tracking system, for display of the primary camera view directly on the VPET client. Latest hardware, like Google Tango, enables users to explore the scene interactively. VPET is released as Open-Source software and seeks active participation from interested individuals, research groups, and companies.

Light capture and control

Creative control and understanding the virtual environment requires connection and harmonisation of real and virtual elements with easy-to-use interfaces and controls. Dreamspace has addressed the harmonisation of real and virtual light to assess the impact of physical lights in the virtual environment and to control the physical lighting to match a virtual configuration.

Real light sources are captured with a high-dynamic range (HDR) light-probing device. We use a tracked camera, equipped with a fisheye lens (Figure 6 left), to capture lights from several positions, typically taken on a sampling grid (e.g. one probe every 1m). From these samples, our method can automatically compute the position and intensity of point lights and direction and fall-off characteristics of spot light sources (3). The computed light parameters are communicated to the Dreamspace renderer, so the virtual scene, can be rendered with the same lighting as the real studio lamps (Figure 6, right). These estimated light sources can be edited in the virtual scene, using the VPET editing tools, and the changes are then sent back to the real lamps via a DMX controller interface to update the studio lighting. This gives creative professionals direct control and feedback of the combined real and virtual lighting and the possibility to try to edit the lights and make creative decisions directly on-set.

Filmed immersive spaces

Constructing virtual environments is a high-cost process, requiring teams of artists, to create highly realistic digital assets. Dreamspace has addressed the low-cost capture of real environments as virtual sets, using an omni-directional camera rig and free-viewpoint rendering to support novel viewpoints. Dreamspace has also developed techniques to capture the geometry of a physical space to act as a display surface for multiple projectors. This allows the projection of filmed environments in novel spaces to create a true holographic display, with no head mounted display.

A high-resolution modular camera system has been constructed to support either 360° capture of real environments or panoramic capture for a 180° view in a defined direction. New techniques have been developed, based on reconstructed scene depth to create a single seamless video back-plate from multiple views, and to synthesise novel viewpoints to support free-viewpoint visualisation (5). The calibration of a physical space as a display surface is based on the projection and 360° capture of a calibration pattern in a target space. The calibration step defines the warping required for a multi-projector system to produce a seamless view of a film environment for a spectator at different positions in space.

Creative virtual productions

The Dreamspace project focuses on making the creation and manipulation of virtual content simpler and easier, so it becomes a more creative process. A *LiveView* visualisation system has been developed that connects to the conventional film pipeline, with real-time tracking, live compositing with depth, real-time global illumination rendering, view-dependent video, and intuitive tools for on-set control. These technologies have been tested and validated in real contexts through creative productions that explore the application in film, installation, and performance arts. This section describes a few productions explored in Dreamspace.

Skywriters production

Skywriters is a documentary film of a family business for sky advertisement. The project was a collaboration with the Institute of Animation at Filmakademie Baden-Württemberg and made use of the *LiveView* system for shot planning, visualisation, and interactive manipulation of animation. The director had no previous experience with virtual production technology. After a short introduction, he could direct digital assets and animation, using a VPET tablet in agreement with the director of photography to design the shots. Figure 8 shows the *LiveView* display as part of the interactive shot planning session. The biggest bottleneck found was the scene preparation for the film pipeline. The setup was evaluated as ‘very intuitive’ and with a high potential to increase creativity.

Immersive display

One of the ambitions in Dreamspace is to immerse the viewer inside a space, with no head-mounted display. This was first demonstrated at CVMP 2015. The *LiveView* system provided a virtual window (Figure 9) onto a captured environment for a tracked viewer, with virtual elements that could be controlled using the VPET tools. The environment was captured, using the modular camera array, with offline rendering of novel views for real-time playback. The concept worked well to adjust shot framing for film production interactively, but additional cues are needed to create the illusion of immersion such as a zero latency stereoscopic display.

Figure 6 *Light probing device (left), image of reference object (middle) and rendered image of the object with estimated lighting (right)*

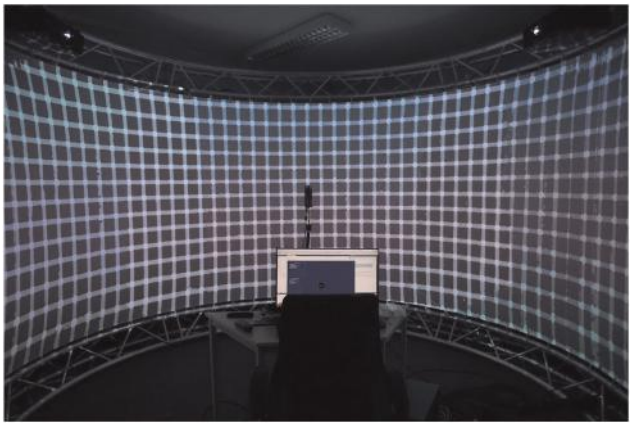


Figure 7 *Calibration of a multi-projector system to create a seamless immersive 360 view*

Figure 9 *Viewer's perspective*

Figure 8 *Virtual Production Skywriters, CGI final shot*

Absence performance

Another Dreamspace ambition is to research the possibilities of applying Virtual Production technologies in live performances, to tell stories in new ways, create new experiences, and even to allow audiences to become part of the story. In *Absence*, a virtual environment was brought on stage using the Dreamspace multi-projector system to manipulate reality. A blend of real and virtual objects was created by projecting virtual content onto both the physical set and the performers, who were tracked using a motion capture system. The concept worked well, since the illusion of 'reality' was supported by the use of real-world objects and an accompanying narration.

Figure 10 '*Absence*' *with set reconstruction, virtual texturing, lighting, augmented transparency, and augmented performers tracked using real-time motion capture suits*

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 30–35

Conclusions

Virtual Production, today, is utilizing more rapidly emerging technologies and new workflows for greater creativity and efficiency. There are two constant pressures, the financial demand for more cost-effective productions and creative demand for greater flexibility and assessment on-set. In addition is the demand from audiences for more advanced story telling. In the creative arts, the demand is for tools and technologies that enhance creativity and experimentation, but at a cost that makes these techniques accessible in a wide range of contexts.

The Dreamspace project has developed key innovations that advance the state-of-the-art in conventional virtual production: real-time depth capture for natural environments, heterogeneous computing for real-time compositing, distributed real-time global illumination for high-quality rendering, capture of physical lights to harmonize real and virtual elements, capture of physical spaces to create projected immersive environments, and intuitive tools to control digital assets connected to conventional film pipelines. These innovations have focused on making these techniques more accessible to a range of budgets, removing the need for a dedicated studio or a green screen shoot, automatically delivering final quality tracking, compositing, and rendering and integrating with live performance environments. The hope that these tools will artistically, practically, and financially open up virtual production techniques to a range of productions and immersive experiences.

Acknowledgments

This work is in part funded from the European Commission's Seventh Framework Programme under grant agreement no 610005 (Dreamspace project).

This work had major contributions from: (The Foundry:) Adam Cherbetji, Guillaume Gales, Marcelo Maes, Alan Purvis, (UdS:) Farshad Einabadi, Richard Membarth, Arsène Pérard-Gayot, Georg Tamm, Jonas Trottnow, (CREW:) Vicky Vermoezen, Koen Goossens (Filmakademie:) Simon Spielmann, Andreas Schuster, Kai Götz (iMinds:) Patrik Goorts, Vincent Jacobs, Lode Jorissen, Steven Maesen, and Sammy Rogmans.

References

1 Dreamspace web page: http://www.dreamspaceproject.eu/
2 Trottnow, J., Götz, K., Seibert, S., Spielmann, S., Helzle, V., Einabadi, F., Sielaff, C. K.H., Grau, O.: 'Intuitive virtual production tools for set and light editing'. Proc. 12th European Conf. on Visual Media Production (CVMP), 2015
3 Einabadi, F., Grau, O.: 'Discrete light source estimation from light probes for photorealistic rendering'. Proc. British Machine Vision Conf. (BMVC), 2015
4 Leißa, R., Boesche, K., Hack, S., Membarth, R., Slusallek, P.: 'Shallow embedding of dsls via online partial evaluation'. Proc. 14th Int. Conf. on Generative Programming: Concepts & Experiences (GPCE), Pittsburgh, PA, USA, 26-27 October 2015, pp. 11–20
5 Jorissen, L., Goorts, P., Rogmans, S., Lafruit, G., Bekaert, P.: 'Multi-camera epipolar plane image feature detection for robust view synthesis'. Proc. 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Lisbon, Portugal, 2015

High dynamic range subjective testing

M. E. Nilsson, B. Allan

British Telecommunications plc, UK

Abstract: This paper describes a set of subjective tests the authors have carried out to assess the end user perception of video, encoded with high dynamic range technology, when viewed in a typical home environment.

Viewers scored individual single clips of content, presented in High Definition (HD) and Ultra High Definition (UHD), in Standard Dynamic Range (SDR), and in High Dynamic Range (HDR), using both the Perceptual Quantiser (PQ) and Hybrid Log Gamma (HLG) transfer characteristics, and presented in SDR as the backwards compatible rendering of the HLG representation.

The quality of HD SDR was improved approximately equal amounts by either increasing the dynamic range or increasing the resolution to UHD. A further smaller increase in quality was observed in the Mean Opinion Scores of the viewers by increasing both the dynamic range and the resolution, but this was not statistically significant.

Introduction

UHD televisions are now retailing in significant numbers, and UHD services are appearing in the market. But while these services offer higher resolution than HD services, further improvement could be made to provide an even better viewing experience.

The next improvement in viewing experience is likely to come from a higher dynamic range for video. Consumer televisions are already shipping with much higher brightness and much higher dynamic range than televisions of only two years ago, and non-consumer displays are capable of much higher brightness still. Standards bodies around the world are debating how high dynamic range should be supported from content capture, through broadcast and distribution channels, to end users on television screens.

In this paper, we report the methodology and results of a set of subjective tests to determine how viewers perceive high dynamic range content on a current high-end consumer television, for what we considered typical content, mostly shot outdoors in sunny conditions in the UK. We wanted to quantify the benefit of adopting new technological solutions that support higher dynamic range for the delivery of content services to current high-end consumer televisions.

We also wanted to compare two non-linear transfer functions that have been standardised to support high dynamic range video, the Perceptual Quantiser (PQ), as defined in SMPTE ST 2084 (1), and Hybrid Log Gamma (HLG), as defined in ARIB STD-B67 (2). We also wanted to quantify the effectiveness of the implicit backward compatibility of HLG, with the quality of standard dynamic range delivers to current high-end consumer televisions.

Test content

BT Sport, with support from BBC and Arri, captured content during an America's Cup World Series event in Portsmouth, UK, 23-26 July 2015, in UHD resolution at 50 frames per second with BT.709 colour primaries (3), using Arri Alexa Mini and Arri Amira cameras. We reviewed the many hours of content captured and selected ten test clips of ten seconds duration for subjective testing, as shown in Figure 1. These clips are varied, including one indoor scene and one outdoor night-time scene, but are dominated by scenes with bright sunshine and water. We feel these scenes represent content broadcast during coverage of an event, like the America's Cup.

Processing of test content

The image processing suite, DaVinci Resolve, was used to reverse the LogC transfer characteristics applied in the camera during capture, outputting EXR files at UHD resolution, at 50 frames per second, with linear light RGB samples in half float format, with RGB samples relative to the BT.709 colour primaries.

We developed software to convert these source EXR images to TIFF format, first applying a matrix to map the samples to BT.2020 primaries (5), then applying a single power function, 'gamma', to each sample of each component, then applying a linear scaling factor, and finally, applying a non-linear transfer function. The equation below shows the part of this mapping for the red component expressed relative to BT.2020 primaries, from linear sample R to non-linear sample R', using scaling factor s, exponent γ (hereafter gamma), and an Opto-Electrical Transfer Function (OETF).

$$R' = OETF(s \times R^{\gamma})$$

For SDR, we selected the inverse of the Electro-Optical Transfer Function (EOTF) specified in BT.1886 as the OETF. For PQ, we selected the Inverse EOTF specified in SMPTE ST 2084. For HLG, we used the concatenation of an inverse OOTF and the OETF specified in ARIB STD B-67 as the OETF, where the inverse OOTF comprised scaling the colour components by a factor dependent on the luminance, L, as in the equation below, where R_{scaled} would then be subject to the OETF. The value of 1.2 was chosen as the peak brightness of the television below 1000cd/m^2, and the contrast and gain were determined for black level zero and white level of 800cd/m^2, as specified in BT.2100-0 (6).

$$R_{scaled} = \frac{L^{\left(\frac{1}{1.2} - 1\right)} \times (s \times R^{\gamma}) - Contrast}{Gain}$$

We adopted this methodology, as we had imagined we could choose suitable values of gamma by viewing still images displayed on a Sim2 monitor with peak brightness. However, this proved not to be possible, as the selected values of gamma did not result in good quality video clips when played on the television. Hence, we adopted the approach described below.

The TIFF images produced by the process above were encoded at HD resolution (1920x1080 at 50fps) as ten second video sequences,

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 36–42

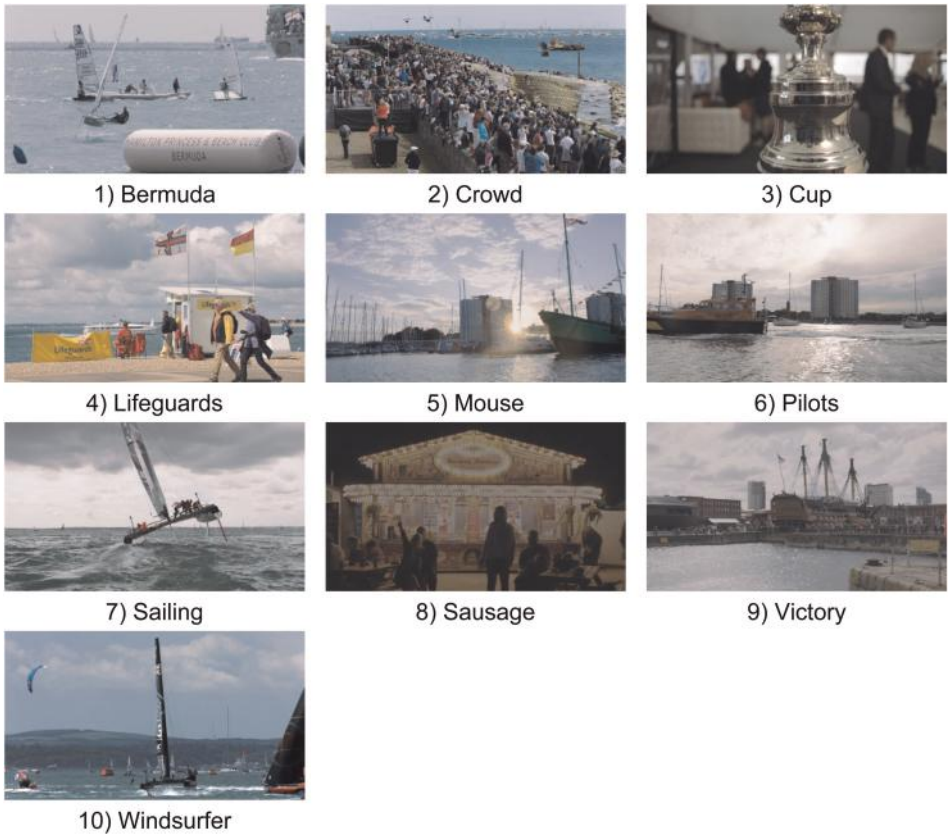


Figure 1 *Single low resolution still images representative of the ten test clips, using BT.709 colour primaries and BT.709/BT.1886 transfer characteristics (4)*

using an Ateme Titan File Encoder to generate HEVC (7) compressed video streams at a bit rate of 30MBit/s within an MP4 file. MP4Box was used to extract the raw HEVC streams, which we processed with our own software to modify the signalling. SDR streams were signalled in the VUI as having transfer_characteristics equal to 1. PQ streams were signalled in the VUI as having transfer_characteristics equal to 16. Two versions of each HLG stream were produced, both having VUI signalling transfer_characteristics equal to 1, but one of them also having periodic repetitions of the alternative_transfer_characteristics SEI message, indicating preferred_transfer_characteristics equal to 18. We could generate one encoded stream for HLG and signal it, in one case, as HDR and, in another case, as backwards compatible SDR.

The resulting modified HEVC streams were multiplexed with an arbitrary audio clip into an MP4 file using MP4 box. The audio was never presented to the viewers and was included solely to prevent the display of the message "audio format not supported."

The values of gamma and scaling factor were chosen for each of the ten test clips and the training clips, for each non-linear transfer characteristic, using a time-consuming iterative process, where we tried different values until we were satisfied with the quality of the clip on the television screen. To get good quality HDR, we found, for many clips, we needed to use different values of gamma and scale factor for HLG, compared to PQ.

We also found the quality of the backwards compatible HLG SDR representation could be improved by choosing values of gamma and scaling factor different to those that produced an optimal HLG HDR representation. We ultimately produced three encoded representations of each test clip for HLG: one, for which we use the term HDR-focussed, which provided a good HDR representation, as close to the PQ representation as possible; a second, for which we use the term SDR-focussed, which provided a good backwards compatible SDR representation, as close to the SDR representation as possible, and a third, for which we use the term Balanced, which provided a reasonable balance between the HDR and backwards compatible SDR representations.

We found, when the HLG HDR quality was high, the backwards compatible SDR representation was often bright and had low contrast. To get a better backwards compatible SDR representation, we frequently had to increase the value of gamma and reduce the scaling factor, but this often had the effect of making the darker parts of the HDR representation too dark.

Although the selection of the content preparation parameters, scaling factor s, and exponent γ, was made using content encoded at High Definition for speed, these eight representations of each clip were generated and encoded at Ultra High Definition (3840 x 2160 at 50fps), using an Ateme Titan File Encoder to generate HEVC compressed video streams at a bit rate of 30MBit/s in the subjective tests.

We also produced two representations in High Definition, besides the above eight in Ultra High Definition, one being a down-sampled version of the PQ representation, and the other being a down-sampled version of the SDR representation but converted from BT.2020 primaries to BT.709 primaries. The PQ representation was encoded with HEVC, using the same encoder, but at a video bit rate of 8MBit/s, and the SDR representation, which was also progressively scanned at 50fps, was encoded with H.264, using the same encoder at the same video bit rate of 8MBit/s, to represent current HD services.

Characteristics of the test content

Statistics were gathered while preparing the test content. The histogram of pixel luminance values was collected for a single image from each test clip for each representation. These histograms are different for SDR, PQ, and the three HLG variants, as different values of gamma and scale factor were selected.

Table 1 shows, for the PQ representation, these values of gamma and pixel luminance statistics, indicating the mean pixel luminance, the lowest and highest pixel luminance values, the 2.5% and 97.5% percentile pixel luminance values, and the corresponding two dynamic ranges. We report these two measures of dynamic range,

Figure 2 *Room configuration for subjective quality evaluation*

as one corresponds to the absolute maximum range, but is subject to extreme individual pixel values, whereas the other gives a range, containing 95% of the pixel luminance values. The table intentionally does not indicate the units in which the luminance is measured, as the values are simply numerical values at the input to the PQ Inverse EOTF function. They are nominally in cd/m^2, but it would be misleading to suggest these were precise values displayed on the television.

Subjective test methodology

The subjective quality evaluation was performed at Adastral Park, in a room with controlled lighting, but not otherwise specifically designed for subjective testing. The aim was to replicate a home viewing environment, as closely as we could, in a workplace room. The background room illumination was set to be about 20 lux.

Test content was presented on a Samsung 65" JS9500 Curved LCD TV, with test clips played back continuously and automatically from USB storage.

Two viewers, separated by a partition, viewed and scored the test content simultaneously. They were seated about 2.6m from the television, a distance found by the BBC to be the median absolute TV viewing distance in the UK in a survey carried out in 2014 (8).

We used the Absolute Category Rating method as specified in ITU-T Recommendation P.910 (9). This is a single stimulus category judgment method, intended for multimedia applications, where the test sequences are presented one at a time and are rated independently on a category scale. After each clip is presented, the viewers are asked to evaluate the quality of the clip shown.

The time pattern for the stimulus presentation is shown in Figure 3. Each ten second test clip is preceded by a seven second period during the middle three seconds of which the clip number is presented. This duration preceding each clip was chosen, because the software running on the Samsung One Connect box causes a banner comprising the filename and a progress bar to be displayed at the start of each clip, and we considered it essential this disappeared, at least, one second before the start of the actual video clip, not at the same time the clip number disappeared. Following presentation of each clip, after one second of black screen, text asking the viewer to vote on the clip was presented for two seconds. The test unit was, therefore, 20 seconds.

Viewers scored each clip independently on the nine-level scale shown in Figure 4.

Viewers were shown four training clips, being four additional clips from the content captured at the America's Cup World Series event, represented with gamma, scaling, and OETF combinations that achieved video qualities representative of the range of qualities seen during the test.

For each test clip, there were ten representations. Eight of these were at UHD resolution (3840 x 2160 at 50fps): SDR, PQ, and for HDR-focussed, SDR-focussed, and Balanced HLG, the HDR representation, and the backwards compatible SDR representation. Two were at HD resolution (1920x1080 at 50fps): SDR using BT.709 colour primaries and PQ.

As ten test clips were used, there were 100 clips to be scored by the viewers. This was too many for each viewer to score each clip. We divided the 100 clips into three groups of 33 or 34, with each group containing six or seven presentations of each source content and six or seven presentations of each encoding format. We created three playlists, termed A, B, and C, each comprising two of these three groups. A group included in the first half of one playlist was included in the second half of another playlist. The ordering of clips within a group was different in each playlist in which it occurred.

Table 1 Statistics of a representative image from each test clip prior to PQ Inverse EOTF

Clip Name	Gamma	Mean Luminance	Minimum Luminance	2.5% Luminance	97.5% Luminance	Maximum Luminance	Full Dynamic Range	95% Dynamic Range
Bermuda	2.0	152	0.518	15	312	597	1152:1	21:1
Crowd	2.4	65	0.107	2	192	561	5246:1	96:1
Cup	2.2	45	0.342	2	256	608	1777:1	128:1
Lifeguards	2.0	176	0.922	9	354	622	675:1	39:1
Mouse	3.4	143	0.020	2	338	417	20829:1	169:1
Pilots	2.2	152	0.097	2	348	424	4378:1	174:1
Sailing	1.9	142	0.435	16	307	322	740:1	19:1
Sausage	1.2	14	0.379	2	51	58	153:1	26:1
Victory	2.0	88	0.537	4	186	255	475:1	47:1
Windsurfer	6.0	137	0.013	1	259	2721	202757:1	259:1

Figure 3 *Stimulus presentation time pattern*

Figure 4 *Nine-grade numerical quality scale*

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 36–42

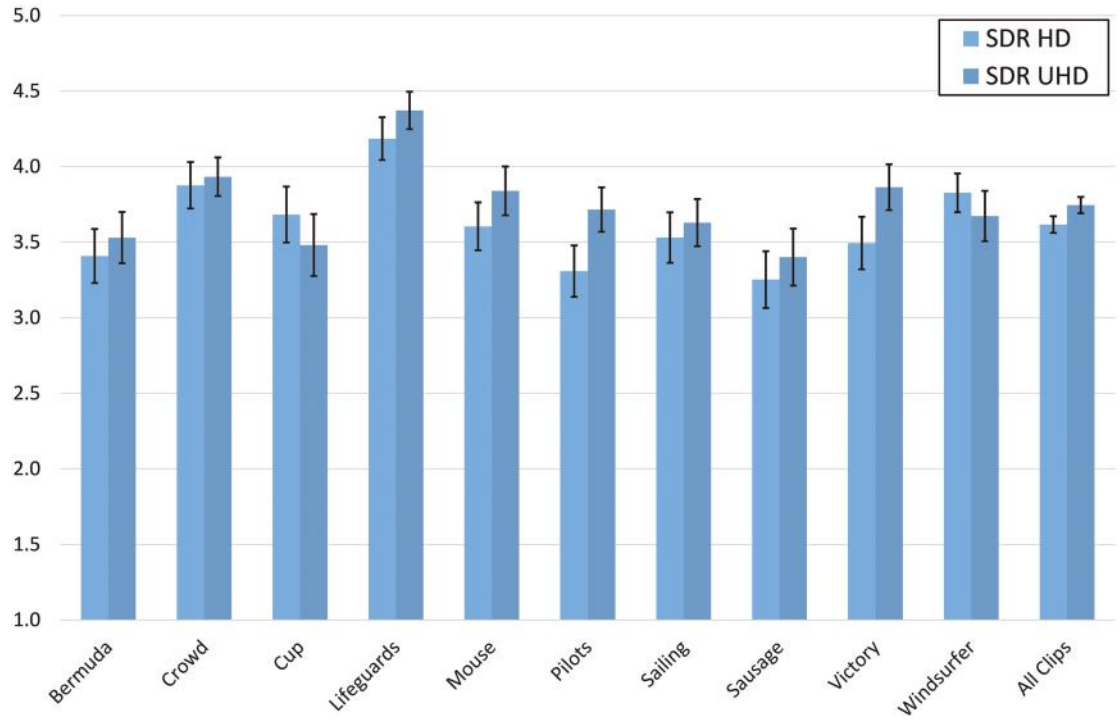


Figure 5 *MOS for each clip in SDR, encoded in HD with H.264 at 8MBit/s, and encoded in UHD with HEVC at 30MBit/s*

All viewers scored 66 or 67 test clips in a test that lasted about 23 minutes, being 67 x 20s, following the period of training.

Approximately the same number of viewers viewed each playlist. Each playlist was structured so the same content was never shown consecutively. Also, the playlists were defined so each test condition (clip and representation) was never preceded by the same test condition in either of the other two playlists.

A total of 122 viewers (94 male and 28 female) participated in the subjective tests, of which 27 considered themselves to be expert viewers. The viewers had an approximately uniform distribution of ages, from 15 to 54, with ten viewers older than 54. Prior to participating in the testing, viewers were screened for visual acuity using a Snellen chart, and for colour blindness, using Ishihara charts. One viewer had 20/30 vision, ten had 20/25 vision, and the remainder had vision 20/20 or better, a large proportion being better. Four viewers were colour blind, but they had above average visual acuity. We chose not to eliminate any viewers based on their eyesight. We applied the outlier identification process described in Annex 2 of ITU-R Recommendation BT.500 (10), despite having many viewers, and found no outliers.

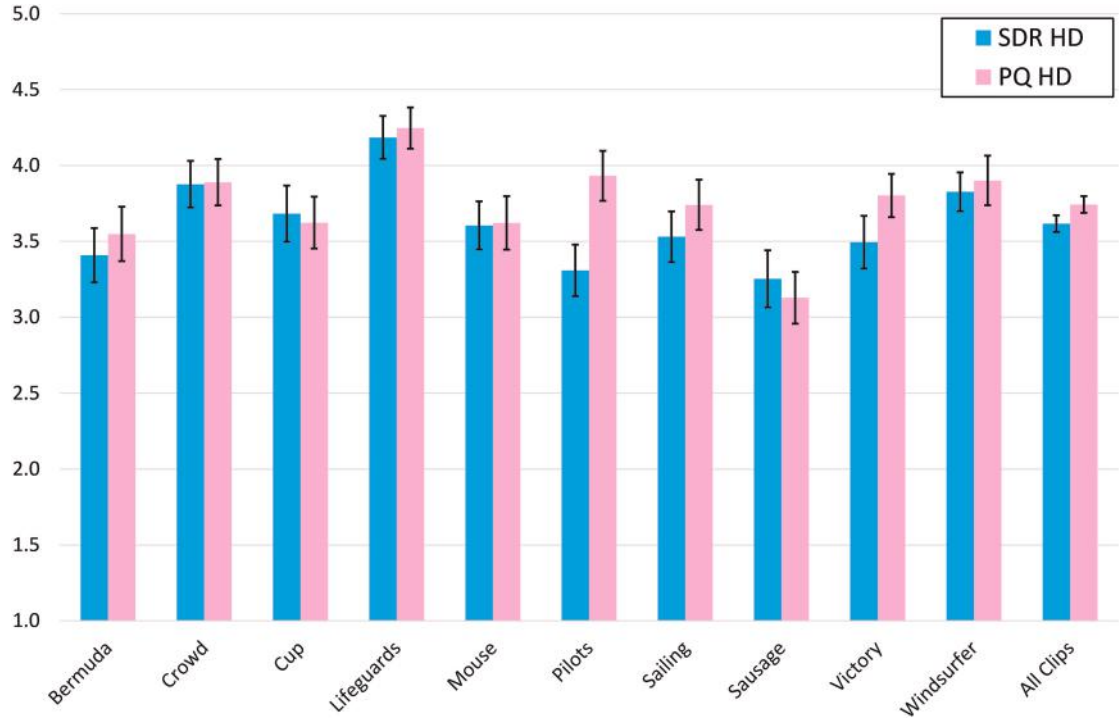


Figure 6 *MOS for each clip in HD, encoded with SDR with H.264 at 8MBit/s, and encoded with HDR using PQ with HEVC at 8MBit/s*

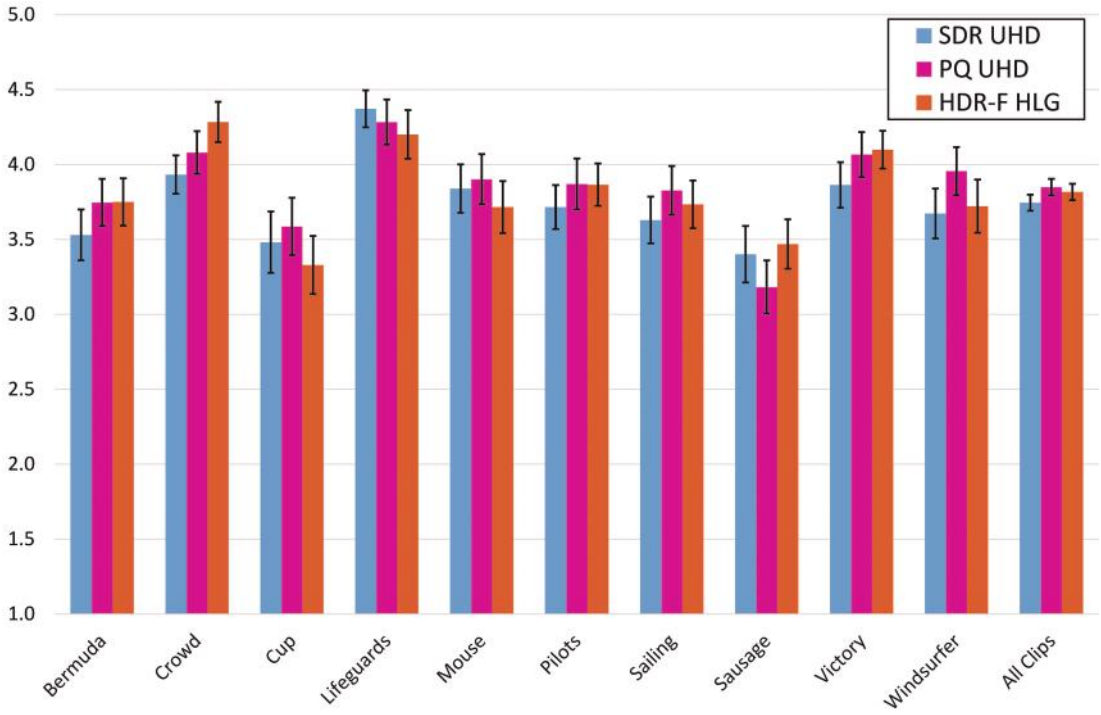


Figure 7 *MOS for each clip in UHD, with SDR, and with HDR using PQ and HDR-Focussed HLG*

Results

The viewers' scores for each clip were mapped to values in the range 1.0 (Bad) to 5.0 (Excellent) in increments of 0.5. These were averaged across all viewers to determine Mean Opinion Scores (MOS) and 95% Confidence Intervals, as specified in Annex 2 of ITU-T Recommendation BT.500.

In the following charts, showing the MOS and Confidence Intervals, we used blue for SDR, pink for PQ, and red for HLG, and lighter shades for HD resolution.

Averaged over all ten clips, UHD resolution was statistically significantly better than HD resolution, with the overall MOS increasing by 0.128 from 3.617 to 3.745.

Eight of the ten clips had higher MOS for UHD, but due to the smaller number of samples and wider confidence intervals, only Pilots and Victory were statistically significantly better than HD.

Averaged over all ten clips at HD resolution, HDR with PQ was statistically significantly better than SDR, with the overall MOS increasing by 0.126 from 3.617 to 3.743.

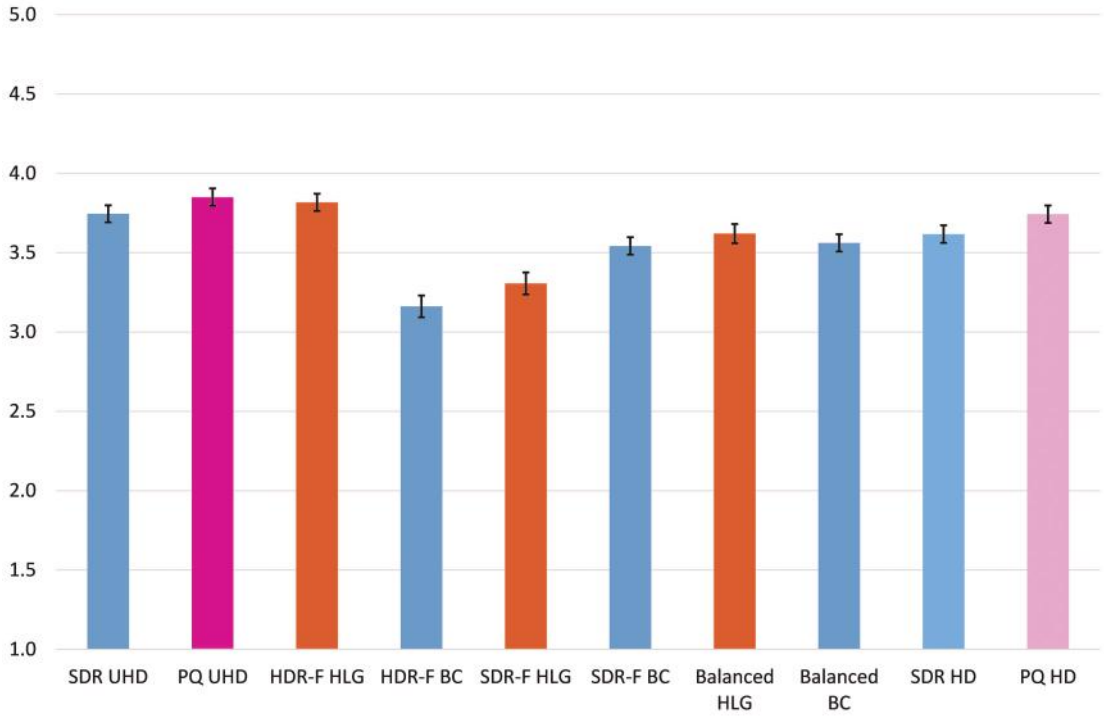


Figure 8 *MOS for each encoded format*

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 36–42

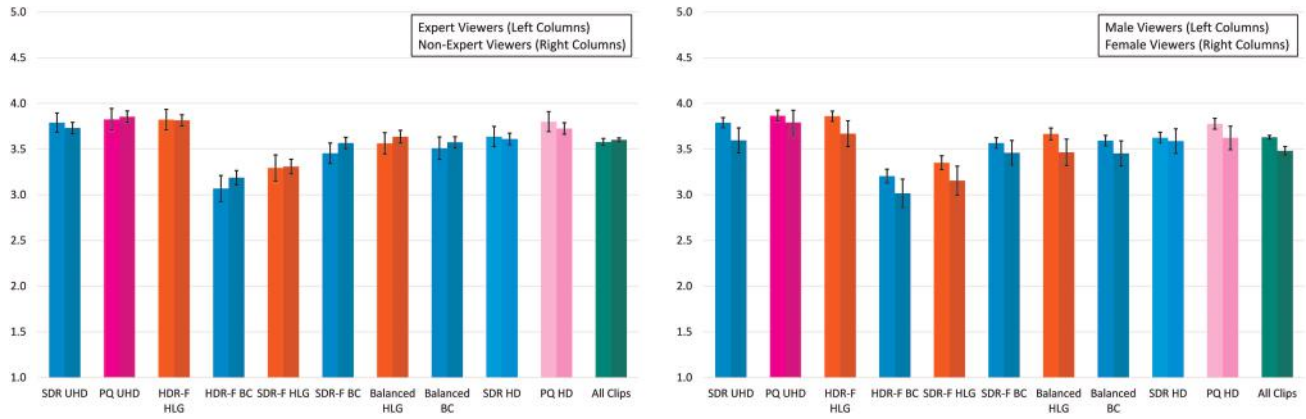


Figure 9 *Overall MOS by viewer expertise (left) and sex (right).*

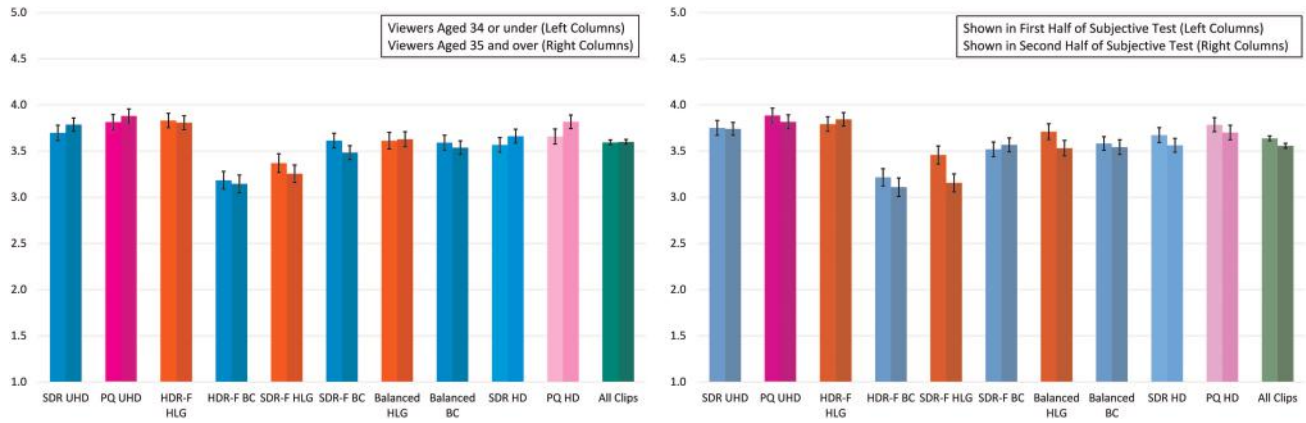


Figure 10 *Overall MOS by viewer age (left) and viewing order (right).*

Eight clips had higher MOS for HDR but due to the smaller number of samples and wider confidence intervals, only one of these, Pilots, was statistically significantly better than SDR.

Averaged over all ten clips, MOS were higher for UHD HDR, with 3.850 for PQ and 3.817 for HDR-Focussed HLG, compared to 3.745 for UHD SDR.

UHD PQ was nearly statistically significantly better than UHD SDR, with the confidence intervals overlapping by only 0.0038. If the one viewer with near zero correlation with average MOS was eliminated, the confidence intervals would have had a gap of 0.000007.

This result was unexpected, as during content preparation, we observed improvements when using HDR, notably in the detail in the clouds and in the sparkle on the water. This was presumably not so noticeable to the viewers when the content was presented in a randomised order. We are not surprised by PQ and HDR-Focussed HLG being statistically indistinguishable, as we considered them to be very similar during content preparation.

The MOS for each of the encoded formats, averaged over the ten clips, are clustered into four groups.

Within each group the MOS are statistically indistinguishable, but they are statistically distinguishable from all encoded formats in the other three groups.

The highest quality group contains the four formats, PQ HD, SDR UHD, PQ UHD and HDR-Focussed HLG, which all outperform SDR at High Definition. The lowest quality group contains the Backwards Compatible representation of the HDR-Focussed HLG, which we considered bright and low in contrast for many clips. The second lowest quality group contains the HDR representation of the SDR-Focussed HLG, which we considered dark for many clips.

The poor performance of the two lowest quality groups did not surprise us, in one respect, as these results are consistent with our opinions formed during content preparation, but surprised us in another, as before starting the project we expected and hoped for better performance from HLG. We consider it an on-going piece of work to gain a better understanding of these issues with HLG.

The similar MOS for Balanced HLG, 3.620 for the HDR and 3.561 for the Backwards Compatible representation, support the opinion we generated a good balance between the HDR and Backwards Compatible representations. However, both are statistically indistinguishable from SDR at HD resolution with MOS equal to 3.617.

Figure 9 shows there was no statistical difference between the 27 expert viewers and the 95 non-expert viewers, with the later scoring higher by only 0.024. It also shows the 28 female viewers' MOS were lower for every format and statistically significantly lower overall, with overall MOS of 3.482, compared to 3.630 for the 94 male viewers.

Figure 10 shows the MOS of the 66 viewers aged 35 or over, 3.599, was statistically indistinguishable from the MOS of the 56 viewers aged 34 or under, 3.593. It also shows MOS for clips, when shown in the first half of the subjective test, 3.636, was statistically significantly higher than the MOS for clips when shown in the second half of the test, 3.556, suggesting viewers became more critical as the test proceeded.

Conclusions

We have carried out a set of subjective tests, using content we believe could be typical of a live outside broadcast event, with many viewers in a 'home-like' environment. The test results suggest the quality of Standard Dynamic Range High Definition services could be improved by approximately equal amounts by either increasing the dynamic range or increasing the resolution to UHD. An additional smaller increase in quality could be achieved by increasing both the dynamic range and the resolution, although this was not statistically

significant. The tests found the Hybrid Log Gamma system achieved approximately equal High Dynamic Range video quality as the Perceptual Quantiser scheme, although the performing its implicit backward compatibility was disappointing. This issue with the Hybrid Log Gamma system requires further study.

Acknowledgements

The authors thank Richard Moreton and his colleagues at Samsung for the loan of a Samsung One Connect box, and for the development of the software for the One Connect box to support the functionality used in the work reported in this paper.

The authors thank Professor Alan Chalmers and his colleagues at the University of Warwick in the UK for the loan of a Sim2 monitor, the provision of a PC player application to use with the Sim2 monitor, and their help and expertise to use this to assist with the preparation of content used in the subjective tests reported in this paper.

References

1 SMPTE ST 2084:2014: 'High dynamic range electro-optical transfer function of mastering reference displays'
2 ARIB STD-B67: 'Essential parameter values for the extended image dynamic range television (EIDRTV) system for programme production'
3 ITU-R Recommendation BT 709-6: 'Parameter values for the HDTV standards for production and international programme exchange'
4 Recommendation ITU-R BT.1886-0: 'Reference electro-optical transfer function for flat panel displays used in HDTV studio production'
5 ITU-R Recommendation BT.2020-2: 'Parameter values for ultra-high definition television systems for production and international programme exchange'
6 Recommendation ITU-R BT.2100-0: 'Image parameter values for high dynamic range television for use in production and international programme exchange'
7 ISO/IEC 23008-2:2015: 'Information technology – High efficiency coding and media delivery in heterogeneous environments – part 2: High efficiency video coding'
8 Noland, K., Truong, L.: 'A survey of UK television viewing conditions', 2015, http://www.bbc.co.uk/rd/publications/whitepaper287
9 ITU-T Recommendation P.910 (04/08): 'Subjective video quality assessment methods for multimedia applications'
10 ITU-R Recommendation BT.500-13: 'Methodology for the subjective assessment of the quality of television pictures'

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 36–42

Directing attention in 360-degree video

Alia Sheikh, Andy Brown, Zillah Watson, Michael Evans

BBC Research and Development, UK

Abstract: 360° video and Virtual Reality are powerful techniques for giving viewers a sense of *'Being There'* [1], and are becoming increasingly popular. However, giving the viewer the freedom to look around also results in a reduced ability for filmmakers to direct the viewer's attention, a serious impediment to successfully telling a story within a 360° environment. We have created a number of 360° clips, filmed in such a way as to demonstrate and test several unobtrusive techniques for directing a viewer's attention within a 360° panorama. We have evaluated these techniques in a user study in which participants viewed these clips using a head-mounted display. Qualitative and quantitative data from these tests have been analysed to evaluate the effectiveness of the different attention-directing techniques. Qualitative data was also captured to explore the effect of the camera being addressed directly, and the viewers' responses to action occurring at a range of distances.

Introduction: visual attention and cinematography

360° video is a special case of virtual reality (VR) in which the audience views a sphere (or near-sphere) of video centred on a single position. 360° formats offer the filmmaker both opportunities and challenges. Unconstrained by a prescribed view, the viewer experiences a video environment in a way that correlates more closely to real life. However, this comes at the cost of limiting the set of techniques open to the director: the use of different camera angles and the ability to cut between them, differential focus and moving camera techniques are all constrained. In conventional TV and film, such techniques can be used by the filmmaker to take the viewer on a specific path through a narrative, ensuring the viewer's attention remains on the elements considered important to the story. In 360° presentation, however, the use of such techniques could have a negative impact on the user's experience, reducing their feeling of control and potentially inducing discomfort. Since some of the key benefits of 360° video are a result of the viewer's control over their own gaze, the filmmaker must allow the viewer to retain that agency and direct gaze using subtler, unobtrusive techniques.

As 360° content is rapidly evolving, directors are developing a new grammar of filmmaking. In addition to accepting a lower level of control over the audience experience, the basic methods for directing attention are starting to be explored, for example by using movement, sound and lighting cues. We seek to understand how effective some of these techniques are through more rigorous audience testing.

Research questions

This paper describes the development and presentation to viewers of some specially produced 360° video material, created to allow us to probe specific directorial mechanisms for directing visual attention in 360° footage, as well as some closely related questions about the subjective experience of this kind of video. Our key research questions are:

1. What attracts attention, what refocuses attention, and what techniques can a filmmaker use to direct the attention of a viewer?
2. How does the distance at which action occurs impact the experience of the viewer?

Are presence, immersion and enjoyment affected by characters in the content addressing the camera directly? We filmed a number of one-take single-shot setups with actors. Each setup was designed to test a specific attention directing technique, or answer a particular research question.

Clips A_1, A_2, A_3 and A_4, were designed to explore directing attention, and were filmed indoors in a large gymnasium. Each clip begins with a clear element of interest, and then uses different methods to try to direct the viewer's attention to a new element of interest introduced later in the shot. Each clip starts with two actors, clearly in view, having a conversation; this was the only action in the scene early on, and lasted at least 45 seconds before any other cues were introduced. Thus, we could be reasonably confident that the viewers would have the opportunity to familiarise themselves with the environment, and that their attention would be drawn to (and ideally retained by) the conversation. Another actor was introduced into an empty portion of the scene (behind the viewer if they were looking at the first two actors), and each clip used a different combination of visual and, in some cases, audio cues to direct the viewer's attention to the newcomer.

Clips D_{2m}, D_{3m}, D_{4m} depicted two actors practising a stagefight. As we wanted to assess the impact of distance on the viewer's comfort-level with the scene, we asked the actors to repeat this sequence at a distance of 2m, 3m and 4m from the camera. Previous studies show that people's comfort at different levels of interpersonal distance is highly context-driven [4], so in this test we filmed an activity that we anticipated people would react consistently to. Testing the reactions of viewers to the presence of people who are being active, but not threatening, allowed us to test for a more physical, instinctive response, as opposed to a more considered one.

In contrast to the other clips, the one used to test presence (P) featured actors explicitly acknowledging the presence of the viewer, with both actors appealing to the viewer to support their side of an argument. This clip is very similar to a clip that was used to acclimatise participants to 360° video, in which the same actors have an argument in the same location, but do not acknowledge the viewer. This allows this direct style of viewer engagement to be investigated.

Research methodology

Participants were recruited through an external agency and a local college. There were 26 participants in total, all of whom took part

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 43–47

Table 1 Cues used to direct attention in clips A_1-A_4

Clip	Summary	Cue 1	Cue 2	Cue 3
A_1	Motion across main characters	Bystander walks to target		
A_2	Motion across main characters with gestural cue	Bystander walks to target, waving		
A_3	Motion across main characters with audio and gestural cues	Target shouts "Alia"	Bystander responds with wave and "Hi"	Bystander walks to target
A_4	Motion of a main character following gestural and audio cues	Main character looks at target	Main characters talk about target	Main character walks to target

in the 'directing attention' part of the study. 17 of the participants also viewed clips designed to understand more subjective aspects of the experience, and their impressions were captured using a questionnaire and a semi-structured interview. In each case, a participant's session lasted under an hour. Video clips were between 20 and 180s in duration and consisted of monoscopic video and stereo audio. Participants viewed the content on a head-mounted display (Oculus Rift) whilst software continuously logged their head orientation within the scene; audio was delivered through a pair of Beyerdynamic DT 770 Pro headphones. Participants were standing while viewing the clips, to match (approximately) the camera height during filming, whilst a researcher supervised for physical safety.

Participants viewed two initial clips for acclimatisation purposes. For virtually all participants this was their first experience of 360° video on a head-mounted display, and certainly under controlled conditions. An initial acclimatisation piece shot in a busy street during the Edinburgh Fringe Festival familiarised them with the style of presentation and their ability to change their orientation to look all around them. A second acclimatisation clip allowed participants to become familiar with the style, actors and location used in a number of the subsequent clips.

After acclimatisation, participants were shown one of the directing attention clips (A_1, A_2, A_3 and A_4), followed by the presence clip (P) and two of the three distance clips (D_{2m}, D_{3m}, D_{4m}). In all cases, head orientation was logged and participants asked for general feedback; for some clips, participants were also asked to rate their levels of enjoyment and sense of immersion, and their ability to follow the action, using a 5-point Likert-style scale.

Directing attention results

Each of the clips involved two main characters, who were having a conversation on a bench, the target, who appeared on the opposite side of the viewer to the main characters, and a bystander. The cues used in each clip to direct attention towards the target are given in Table 1, while Figure 2 illustrates the scene.

Figure 3 gives an overview of the results for each clip. These plots show the percentage of people who had seen the target over the time since the first cue. Comparing clips A_1, A_2 and A_3, it can be seen that whilst motion cues alone have some effect, the addition of audio and gestural cues increases the effectiveness with which we can direct attention.

Figure 1 *A participant viewing the Royal Mile acclimatisation scene in the lab study*

Figure 2 *The scene for the directing attention clips A_1-A_4. The main characters are seated in the centre, the target is on the far left; the bystander has just left her starting position below the clock*

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 43–47

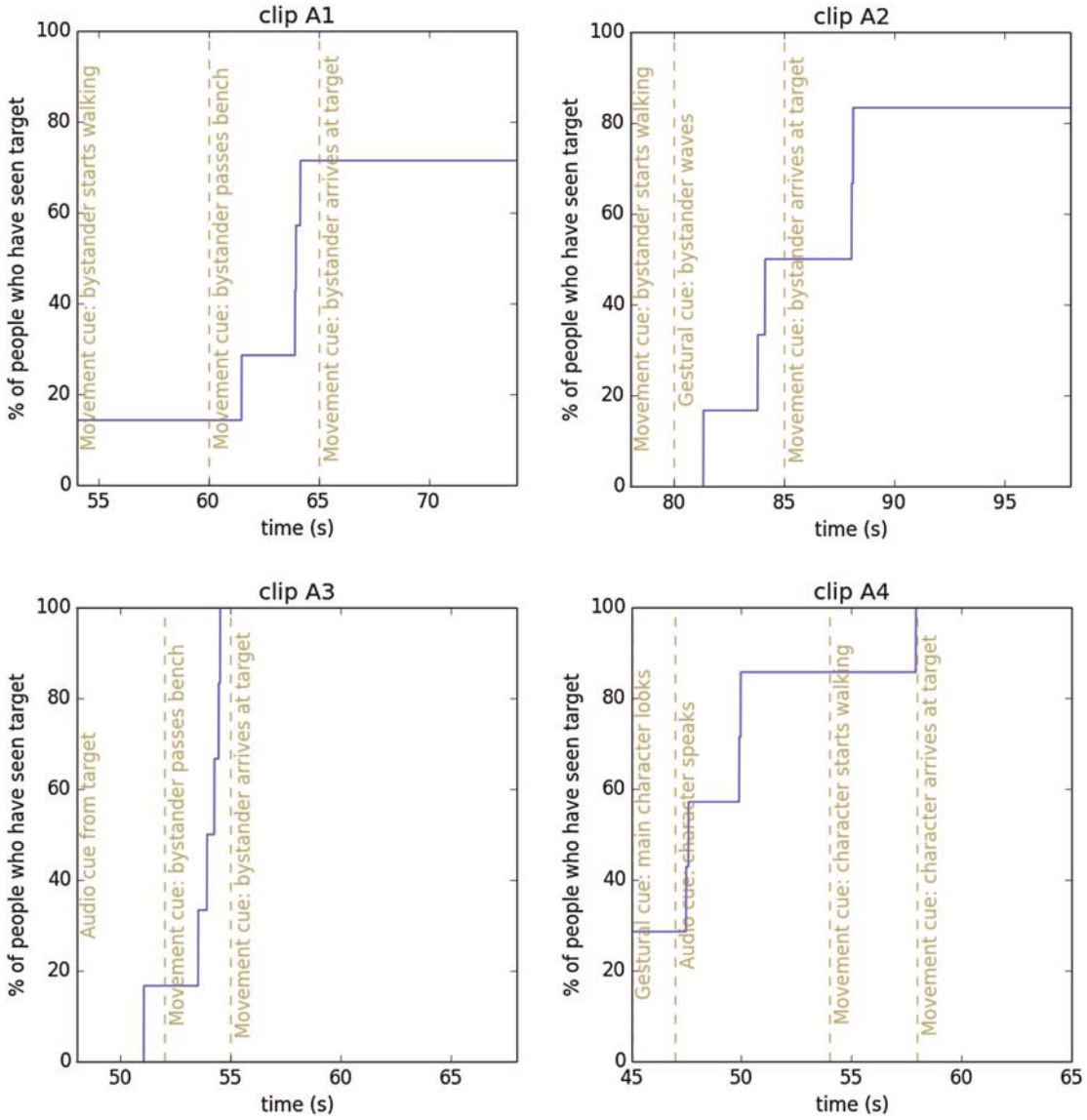


Figure 3 *Plots illustrating the effectiveness of the directing attention cues. These show the cumulative percentage of participants who had seen the target at any time since the first cue (the times of cues are shown with vertical dashed lines)*

The bystander walking past was, in isolation, moderately effective (clip A_1), although 2 of the 7 participants did not see the target. In both cases, the cue was seen, but not followed to the target. Supplementing this with a second visual cue – the wave from the bystander (clip A_2) – alerted the viewers that the bystander was walking towards someone. This was more effective, with only 1 participant from 7 missing the target – this person did not see the wave and did not follow the bystander's motion. Adding an audio cue from both target and bystander meant that some cue was perceived no matter where the viewer was looking, and was much more effective – all viewers saw the target, and attention shifted very quickly (within 7s).

Clip A_4 used different techniques, involving one of the protagonists making reference to, and ultimately moving towards the target. In this case, it can be seen that of the 7 viewers, 2 people were looking at the target before the first cue, 4 responded to the second cue (the mention of the target), while the last person followed the protagonist walking across to the target. It would be interesting to evaluate in more detail whether a cue from a protagonist is more effective than a cue from an inactive bystander.

Table 2 Participant preferences for each pair of distance clips. This reports the number of participants preferring either the closer or the farther of each pair

Comparison	Closer	Further
2m vs. 4m	4	1
3m vs. 4m	4	1
2m vs. 3m	1	5

Subjective experience results

Subjective data were collected from participants for several of the clips. These allow us to explore preferences, and the reasons behind them. In particular, we are interested in how participants felt about watching the fight scene at two distances, which distance they preferred, and why, and how participants felt about the actors addressing the camera directly.

Distance

Participants showed a clear preference for the fight training taking place at 3m. Both the 2m and 3m distances were preferred when compared with 4m, but when 2m and 3m were compared directly, there was a clear preference for 3m (Table 2).

Participant's ratings showed that the distance did not impact their ability to follow the action, but it did affect their enjoyment of the clip, and their sense of immersion. The enjoyment ratings matched the preferences to show that 4m was too distant, but there was no clear difference in the ratings given to the 2m and 3m.

The interviews revealed a number of reasons for this preference. When taking place at 4m, the action was felt to be too distant.

> '*It felt like you were watching something across the street*' *[P13, 4m]*

In contrast, the same participant found the 2m clip much more realistic and immersive:

> '*It didn't feel like I was watching TV or anything, it felt like I was actually there*' *[P13, 2m].*

Other people, however felt that at 2m the action was too close, in the personal space of the viewer:

> *[It was]* '*very, very close to where I was… that wouldn't usually be happening that close to somebody*' *[P01, 2m]*

> '*When they stood too close, it felt like they were more in your personal space*' *[P12, 2m]*

The impact of acknowledging the viewer

Participants were prompted for their thoughts on the presence clip (P) by being asked to rate their experience using the (informally worded) question: "*How immersed did you feel; how much did it feel like you were there? 1 represents* '*not at all*'*; 5* '*very much*' ". Comparing the ratings given to the acclimatisation clip and the Presence clip, the latter receives a higher proportion of ratings of 4 or 5 (94% vs. 47%). The enjoyment ratings followed a similar pattern: for the acclimatisation clip, 56% of participants gave a rating of 4, and 12% of 5; for the Presence clip these values were 50% and 35% respectively. Other than the actors addressing the camera, these clips were similar: both were filmed in the same location, and involved the same actors in an argument. In both cases the actors spent some time moving around the camera. The major differences were that P1 was longer (3 minutes rather than 1.5 minutes), and was always shown second.

While the ratings give an indication of the impact of this technique, the qualitative feedback was much richer. The overwhelming reaction to the actors talking to and gesturing towards the camera was positive:

> '*After a while it felt like I was just standing talking to two friends… it felt like real life to me, not just a staged environment*' *[P13]*

> '*It kind of felt like you were actually involved in the conversation… I thought it was good… it makes you feel like you're there.*' *[P4]*

Interestingly, the sense of immersion was significantly affected by a minor gesture by one of the actors in the D_{4m} distance clip: as the fight finishes, the female actor points to where she is going next — this happened to be close to the camera, and several participants commented that she pointed at them, and that this made them feel more part of the scene. Participants also commented favourably on being looked at by passers by in the Royal Mile acclimatisation clip.

Discussion

What attracts attention, what refocuses attention (how do viewers distribute their visual attention), and what techniques can we use to direct the attention of a viewer?

Four techniques for directing viewer's attention to one portion of a single shot scene were evaluated. The most effective used both audio and visual cues from the target area and the part of the scene where the viewers were assumed to be looking. The most unobtrusive technique was using a bystander to walk across the action towards the target; this was effective for 5 of the 7 participants who watched the clip. Audio cues have the advantage that no assumption is made about the viewer's focus of attention at the time of the cue. Even without fully spatialised audio, the use of sound also alerts the viewer that there is something to see; with the visual cue alone, participants sometimes followed the cue, but not as far as the target. When both audio and visual cues were used, all participants saw the target.

How does the distance at which action occurs affect the immersion or enjoyment of the viewer?

In the context of viewing the fight training scene in the distance clips D_{2m}, D_{3m} and D_{4m}, there was a general consensus among participants that when the action occurred at 4m it felt too distant, but 2m felt unnaturally close. 3m provided a good balance of being close enough to see clearly and provide a sense of immersion, but far enough that it wasn't happening in their private space. The evidence in the literature from virtual reality [2, 3] suggests that "the response to invasion of virtual personal space shows the same trend as the response to the same stimuli in a live setting" [2], so it is interesting that this experiment indicates that 360° video may have similar effects. In addition, it is known that people under threat maintain a greater personal distance [4], so the comfort felt at a given distance will vary with context (the 1-3m range in the close setting comes into the viewer's personal distance zone [5], where the practice fighting involving large body movements could feel threatening). Thus, the results found in this experiment may reflect the balance of the viewer's desire to maintain a 'safe' distance from the action with their ability to have a good view of the action, and feel part of it.

Furthermore, it is known that depth perception is distorted when viewing static camera monoscopic 360° video, with objects appearing further away than they are, an effect that scales approximately linearly with actual distance [6]. The material used in this experiment was filmed using a static camera and in an environment (a relatively empty rooftop courtyard) with few depth cues (e.g., occlusion). These combine to create an apparent 'collapse in perspective' where, as the actors move further from the camera they become harder to separate from the background, and thus feel artificially further away. Compromised depth perception is a fundamental challenge for monoscopic 360° video that needs to be considered by directors, but it can be mitigated to some extent by set design or camera motion.

Finally, it should be noted that it will not always be the aim of the filmmaker to keep the viewer at a comfortable distance, but understanding what this is will allow them to manipulate this variable in order to achieve a desired response.

Are presence, immersion and enjoyment affected by characters in the content addressing the camera directly?

There was a strong feeling that participants felt more immersed in the content, and enjoyed it more when they were acknowledged by characters in the scene. The technique of having the actors directly reference the camera as another character was effective, but even pointing at or making eye-contact with the camera without verbal reference had a notable effect.

Conclusions and further work

These experiments have demonstrated the effectiveness of various visual and audio cues for directing the viewer's attention within a 360° scene. We found that the combination of audio and visual cues is more powerful than visual cues alone, this is mainly because audio cues are less dependent on the focus of attention at the time of the cue. While they cannot guarantee that attention will be given to the desired part of the scene, such cues can be used by filmmakers to guide the audience through a narrative. We also found that participants reacted positively to being directly addressed or acknowledged by characters within the scene. The response of participants to fight practice occurring at different distances in 360° video matched what would be expected in

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 43–47

real-world scenarios and virtual environments. Thus, we anticipate that camera distance can be used by directors to induce particular emotions in a way that maps real life. Further research is required to understand how distance is perceived in 360° video, given any specific projection and mapping, and also to understand how other scenarios would be experienced.

We believe that developing an understanding of the user experience of 360° video, through studies such as this, can inform filmmakers and accelerate the development of the craft. These experiments represent only a start, however. There are other techniques that are being used to direct attention, which we would like to explore further. For example: Which lighting techniques are most effective? When rendering 360° video in a virtual environment, how can additional objects be composited on the scene to guide the viewer? Spatial audio is recognised as enabling a richer experience [7]; given that stereo was effective, how much better is fully spatialised sound for directing attention? Other techniques using choreography are also possible; we would like to explore how storytellers versed in the theatre space approach and solve these problems – the theatre is, after all, a single set with a fixed audience viewpoint. What additional blocking techniques used in the theatre can be applied to 360° video?

Furthermore, this is early work using limited numbers of participants: it will be necessary to explore these questions further, moving beyond bespoke test material, to understand how they work in longer sequences with a more defined/driven narrative. This will include researching methods to successfully edit sequences together within a narrative, in ways that feel natural and un-prescribed. Retaining the viewer agency afforded by 360° video is crucial to the experience, so we need a suite of techniques that will allow us to move the viewer through the story without them being aware of the guiding hand of the director.

Acknowledgments

The authors would like to thank Mike Armstrong, Angela McArthur and Vanessa Pope for their assistance with background research, Lianne Kerlin, Mia Roscoe and Maxine Glancy for helping to run the study, and the participants for their time and feedback.

References

1 Cummings, J.J., Bailenson, J.N.: 'How immersive is enough? A meta-analysis of the effect of immersive technology on user presence', *Media Psychol.*, in press
2 Wilcox, L.M., Allison, S.E., Elfassy, S., Grelik, C.: 'Personal space in virtual reality', *ACM Trans. Appl. Percep.*, 2006, **3**, (4), pp. 421–428. DOI: http://dx.doi.org/10.1145/1190036.1190041
3 Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: 'Interpersonal distance in immersive virtual environments', *Pers Soc Psychol Bul.*, 2003, **29**, (7), pp. 819–833. DOI: http://dx.doi.org/10.1177/0146167203253270
4 Uzzell, D., Horne, N.: 'The influence of biological sex, sexuality and gender role on interpersonal distance', *British J. Social Psychol.*, 2006, **45**, pp. 579–597
5 Hall, E.T.: 'A system for the notation of proxemic behavior', *Amer. Anthropol.*, 1963, **65**, (5), pp. 1003–1026
6 Kaufman, L.: 'Sight and mind: an introduction to visual perception', 1974, *OU. Press*
7 Skalski, P., Whitbred, R.: 'Image versus sound: a comparison of formal feature effects on presence and video game enjoyment', *Psychol. J.*, 2010, **8**, (1), pp. 67–84

Towards new forms of news gathering through crowdsourced Live Mobile Streaming systems

Ray van Brandenburg[1], Omar Niamut[1], Arjen Veenhuizen[1], Gert-Jaap Hoekman[2]

[1]*TNO, The Netherlands*
[2]*NU.nl*

Abstract: Crowdsourced Live Mobile Streaming applications, such as Meerkat and Periscope, have seen an explosive growth in their popularity in the past few years. Whereas these applications provide great opportunities for crowdsourcers to directly share their experiences around live events, the unedited nature of these user-generated video streams makes them less suited for enriching news broadcasts or event reports. For such cases, the ability to select and edit streams as they come in, or to communicate with reporters in the field, are primary requirements for any editorial office or newsroom. This paper reports on the design of, and experimentation with, a crowdsourced live mobile streaming system and application for: requesting, receiving, filtering, directing, editing, and broadcasting live video streams from both consumers and professionals. This enables new forms of crowdsourced news gathering. The paper incorporates results from a number of technology validation tests and demonstrations, performed in collaboration with Dutch media partners.

Introduction

With the growing popularity of social media sites, online video services, and smartphones, content consumers are recording, editing, and broadcasting their own stories. Social media sites have evolved from text and photo streams to a rich medium for sharing audiovisual content. More and more mobile users are capturing and sharing video content, which can largely be attributed to the increased availability of video-recording capabilities on personal and mobile devices such as smartphones and their integration with online content-sharing platforms [1]. Amateur video capturing has also evolved from a personal hobby to prosumer usage for e.g. eSports broadcasting and citizen journalism. New crowdsourced live mobile streaming (CLMS) applications, such as Meerkat[1] and Periscope[2], have seen an explosive growth in popularity in the past few years [2]. In such systems, a large number of geo-distributed users publish live video streams from their mobile devices for an even larger audience to view. As a result, the role of these mobile Internet users during live events sees a shift, from taking part in a traditional passive audience, to acting as a content creator/participator, i.e., a crowdsourcer. In particular for live events, people from around the world are offered the ability to watch what is happening through the "eyes" of the crowdsourcer. CLMS systems offer event organizers a new and rapid way to distribute content to audiences; content that is unpolished, but genuine and real.

Most of the existing CLMS applications provide opportunities for crowdsourcers to directly share experiences around live events, but the unedited nature of user-generated video streams make them less suited for enriching news broadcasts or event reports. Broadcasters and media outlets are on the lookout for the best method to incorporate the potentially valuable source of content into their existing networks and workflow, for augmented TV broadcasts and citizen journalism in contemporary newsgathering. Curation of these crowdsourced streams, in combination with providing context and professional reporting, are crucial aspects of such a method. This comes with a set of functional (i), and technical (ii), challenges, e.g. (i) ensuring a low threshold for crowdsourcers with single-button streaming, filtering the incoming streams based on quality, allowing editors to 'direct' live streamers, making streamers feel appreciated and thereby more likely to continue streaming, and (ii) providing for reliable low-latency video streaming, a scalable backend and easy deployment, seamless switching between audio and video streams and catering for a huge diversity of sources and sensors.

This paper reports on the design of, and experimentation with, Cameraad, a CLMS system and application for requesting, receiving, filtering, directing, editing, and broadcasting live video streams from consumers as well as professionals, enabling new forms of crowdsourced news gathering. The system features a robust and modular system design and a simple user interface for on-the-fly editing and stream selection by the video editor. In particular, we (i) discuss the underlying design principles and use of open-source and web-based technologies such as WebRTC and GStreamer; (ii) show how we have derived and implemented a cloud-based architecture that allows for rapid deployment of the entire ingest, production and editing system; and (iii) present results from a number of technology validation tests and demonstrations, e.g. during the Grand Depart of the 2015 Tour de France in Utrecht, the 2015 Four Days March in Nijmegen, and a 2016 football game between PSV and Atlético Madrid, all taking place in The Netherlands.

Related work

Industry pioneer Twitch.tv[3] allows anyone to broadcast their content to large numbers of viewers, and the primary sources come from game players, from PCs or from other gaming consoles, e.g., Playstation 4 and Xbox One. According to Twitch's Retrospective Report 2013, in just three years, the number of viewers grew from 20 million to 45 million, while the number of unique broadcasters tripled to 900,000 [3]. Youtube Live and Facebook Live [4] are dedicated features of the respective well-known platforms that allow virtually anybody to easily broadcast themselves to a large audience. Pires and Simon [5] did an early study of YouTube Live and Twitch as emerging live streaming services. YouTube Live is

[1]http://meerkatapp.co
[2]www.periscope.tv
[3]www.twitch.tv

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 48–52

functionally similar to Twitch, but with more general content and a variety of different channels. Their log data found that both services offered a choice of live content at all hours of the day, although they did exhibit diurnal and weekly patterns. They found that 30% of Twitch sessions lasted 60-120 minutes, which appeared to be largely driven by the length of the shared gaming activity. Meerkat and Periscope are among the most recent and popular CLMS applications. Periscope allows users to login with their Twitter account; as such, users are then provided with a list of public live-streams that others are currently broadcasting and one click gives one full access to what the "Broadcaster" is seeing in real-time. The live-stream is also accompanied by a viewer count indicating how many other people are co-watching the stream. In addition, viewers can comment and 'heart' the stream for everyone to see. Streams with the most 'hearts' (or, as Periscope has dubbed them, the "Most Loved") end up with the highest rankings in the app, which attracts even more viewers. Periscope, now owned by Twitter, is set apart from Meerkat in two ways. First, it integrates deeply with Twitter, making it easy to create an account and Tweet a link out to the world. Second, Periscope offers a 'Replay' option where users who weren't around during the live stream can go back and watch.

BeFirst[4], developed by LiveU, leverages their live IP contribution technologies for citizen journalism. BeFirst has two parts: a software component, or "SDK" that plugs into an existing mobile application and that provides live video streaming, and the LiveU Central hosted back-end system that allows control of this stream via any web browser. The integration of BeFirst into existing mobile applications adds the ability to stream high-quality live video from any smartphone that has installed the application directly to a back office. In addition, it allows editors to determine which of their app-bearing users are in any specific geographical area, and message them directly. MakeTV[5] is a live video cloud solution for acquiring, curating and distributing incoming video streams from a variety of devices. Similar in its design to Cameraad, it features multiple stream inputs and outputs, communication with reporters, and newsroom and live production workflow integration. Both BeFirst and MakeTV primarily target professional mobile reporters, rather than untrained crowd reporters.

Within the EU FP7 project STEER [6], an application was developed augmenting a live event broadcast with live user-generated video originating from mobile devices. Event participants made live recordings with mobile devices, that could be streamed and watched fully synchronized with the live broadcasts by viewers at home. In addition, an analysis of social network messages on Twitter was performed to retrieve and show the most relevant posts in conjunction with the video. By providing additional event-related social and video content, viewers can enjoy an augmented view of live events.

Cameraad: CLMS for citizen journalism

Cameraad was developed with the aim of supporting newsrooms to efficiently integrate live user-generated content (UGC) streams, captured and contributed live by citizen journalists. This required seamless integration of crowdsourced live streams into the newsroom workflow and editorial process. The system was co-developed by and tested with NU.nl[6], the largest Dutch online news provider. NU.nl reaches 2.7 million people on a daily basis, and has 4.2 million subscriptions to their breaking news services. Dedicated UGC editors oversee the insourcing, verification and curation of crowdsourced content, such as photos and offline videos, support integration of such content into their news feed, and allows their users to contribute and comment on news events. No less than 30% of their photo content in the domestic news section is crowdsourced, and UGC provides for an important source in developing news stories. The Cameraad CLMS offers

[4]http://www.liveu.tv/befirst
[5]https://www.make.tv/
[6]http://www.nu.nl/

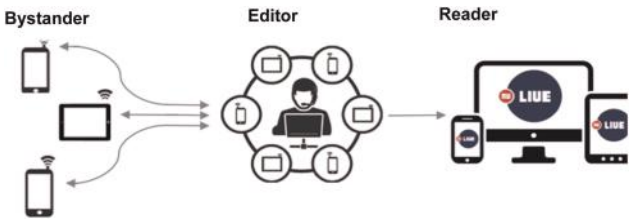


Figure 1 *Cameraad concept*

NU.nl an extension to their crowdsourced inputs, beyond images and offline video, to increase the timeliness of their news reports. In particular, they sought to maximize both the quality and quantity of incoming streams, to filter and select among the incoming streams, and to provision the interaction between crowdsourcers and newsroom editors. Figure 1 depicts the Cameraad concept, implemented as part of the NULive service.

System overview

The Cameraad system, shown in Figure 2, has a modular setup; on the client side, the Cameraad library can be integrated with an existing mobile application for live video streaming functionality; on the server side, a management server and editor web panel allow for service control and news production, with outlets to websites, online video portals such as YouTube and mobile applications. The underlying platform ensures the ingest and routing of incoming video streams.

Client module and editor panel

The client library has been developed as a standalone library for easy integration in mobile applications. The library incorporates the Google WebRTC implementation to stream Opus audio [7] and VP8 video [8], at a typical total bitrate of 800-1024kbps. WebRTC[7] was developed by the Google Hangouts team to support audio and video technology in browsers, without the need for dedicated plugins. The library further provides for communication with the management server via Websockets for initial session setup. A separate peer-to-peer WebRTC-based audio link allows for communication between a streamer and an editor via the editor web panel. Finally, the WebRTC library offers ICE Negotiation with a STUN/TURN server (defined below) for NAT traversal [9]; a crucial feature for reliable video streaming over mobile networks. ICE (Interactive Connectivity Establishment) is a framework to provide reliable IP set-up and media transport.

The editor panel provides an easy-to-use web-based interface (JavaScript, CSS, HTML5) for editors to monitor incoming streams, and to select and route streams to a live newsfeed in an interactive mosaic-style user interface. Separate mosaic and switcher streams are received via WebRTC from the backend, allowing for simple low-latency streaming to the web admin panel, and Websocket-based communication is maintained with the management server. Editors can setup a separate peer-to-peer audio or text chat link for communication with live streamers. Google Maps API integration provides for signaling the current location of crowdsourcers, allowing editors to perform position-based filtering of streams. Figure 3 above shows the client module and editor panel user interfaces.

Management and backend server

The management server handles the video room for each streamer: monitors the streams that join and leave or experience connection issues, and it spawns stream-specific ingest processes for each client, depending on the stream type the client is sending. It is aware of incoming stream details, ongoing interactions with

[7]https://webrtc.org/

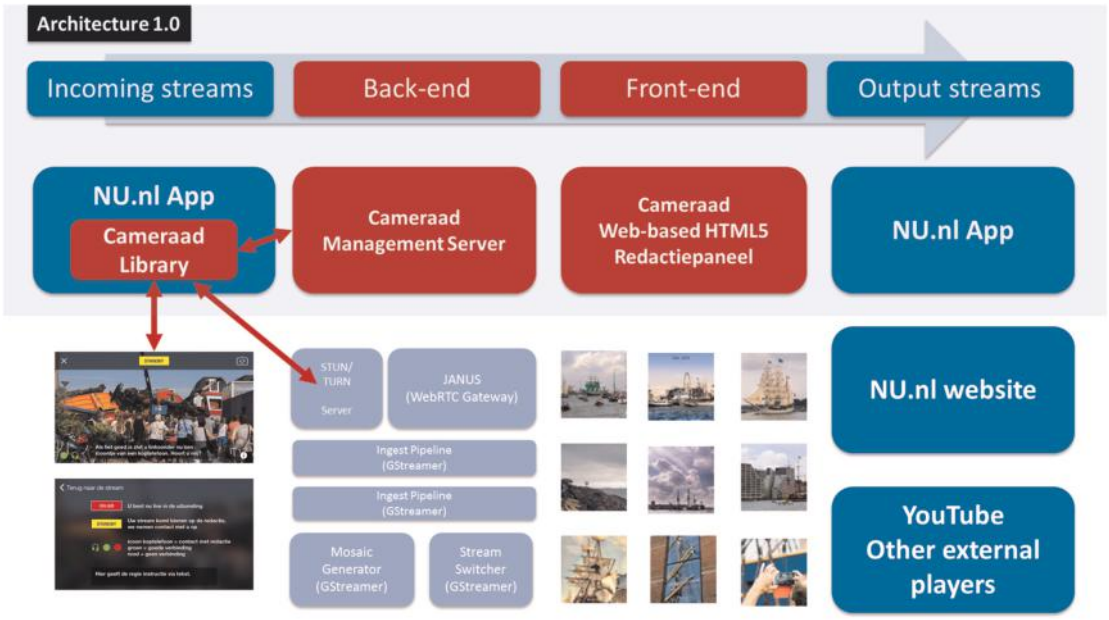


Figure 2 *Cameraad system overview*

Figure 3 *Cameraad client module (left) and editor panel (right) user interfaces*

streamers, and controls stream switching and routing, by signaling the mosaic structure (e.g. size and sources), and which stream to output to which destination (e.g. to the web admin panel or a YouTube Live channel for large-scale distribution). All Cameraad front-end components run on a cloud-based backend platform (see Figure 4), deployed on an Amazon EC2 instance. The backend can scale with the number of incoming streams and has a modular setup, including the following components:

STUN/TURN server; STUN (Session Traversal Utilities for NAT) helps connect IP end-points, by discovering whether they are behind a NAT/firewall, and if so, to determine the public IP address and type of the firewall. STUN uses this information to assist in establishing peer-to-peer IP connectivity. TURN (Traversal Using Relay NAT) addresses this by providing a fallback NAT traversal technique using a media relay server to facilitate media transport between end-points in corporate networks, where STUN is ineffective.

JANUS WebRTC Gateway; provides the means to set up a WebRTC media communication with an application, exchanging messages with it, and relaying RTP/RTCP and messages between applications and the server-side application logic.

All media processing in the Cameraad backend is handled by GStreamer, an open source media framework[8]. The following process can be discerned:

- *Ingest Pipeline*; a custom GStreamer pipeline tailored to receiving a specific type of audio/video feed. Its plugin architecture ensures

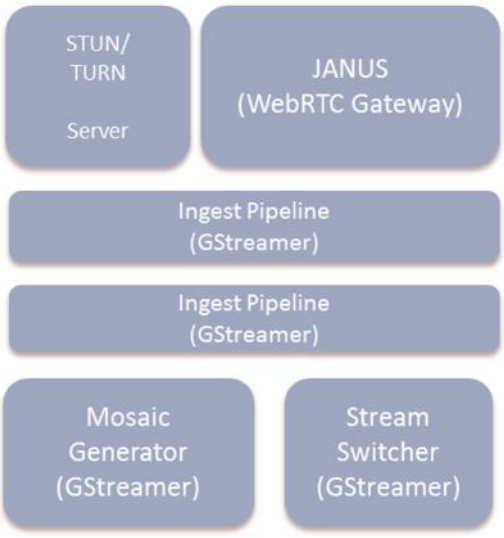


Figure 4 *Cameraad backend*

[8]https://gstreamer.freedesktop.org/

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 48–52

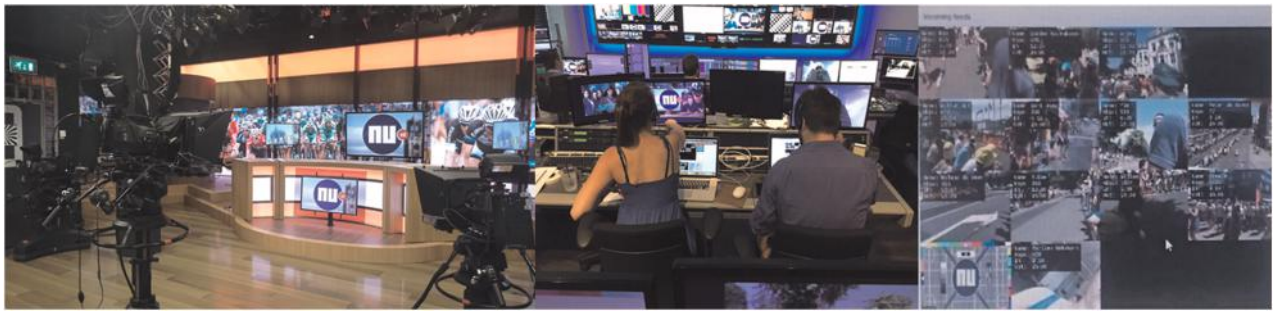


Figure 5 *First large-scale test during Grand Depart Tour de France in Utrecht*

minimal effort to create a plugin to receive a WebRTC stream, a live feed from a drone or a professional broadcast camera. The ingest pipeline output (both audio and video) is decoded in a standard intermediate format which is subsequently used by the *Mosaic Generator* and *Stream Switcher*. Since stream arrivals and departures are ad-hoc, and since the connectivity can vary heavily, the ingest pipeline makes sure that any lost packets and/or broken audio/video frames are replaced and/or duplicated.

- *Mosaic Generator;* dynamically creates a mosaic of zero or more incoming feeds. All tiles in the mosaic have equal dimensions. The output of the mosaic is encoded in VP8 and does not contain any audio feed.
- *Stream Switcher;* facilitates seamless switching between available streams. This is achieved by switching in the uncompressed audio and video domain. The audio/video output is then encoded with appropriate encoders. For WebRTC output, for example, VP8 and OPUS are used. For YouTube, H264 and AAC are selected.

Experiments and trials

The Cameraad CLMS platform and application was tested during several large-scale events. Initial tests during development were performed during Liberation Day, on May 5th, 2015 and also an emerging news story around a fire in a government building. In the first test, editors of Nu.nl reported live on several festival locations around the country, with the resulting live stream visible on the Nu.nl main website. In the second test, the unexpected live report rapidly attracted up to 15,000 viewers, and the live newsfeed was embedded on several other large Dutch news websites.

2015 Grand Depart Tour de France in Utrecht

A first large-scale live test was performed during the Grand Depart, the start of the Tour de France 2015, which took place in Utrecht, The Netherlands. Participants had to pre-register by mail, resulting in 40 semi-friendly crowdsourcers. No professional reporters were included. Incoming and selected streams were incorporated in the newsfeed, produced in a dedicated broadcast studio, and commented on by a professional reporting team. Having the geo-location of streamers available helped in briefing and preparing live streamers prior to their live contribution to the newsfeed. With close to 250,000 bystanders, the mobile network was severely overloaded, and the WebRTC Gateway was unstable. We noted that the value of UGC, to enrich an event that is already

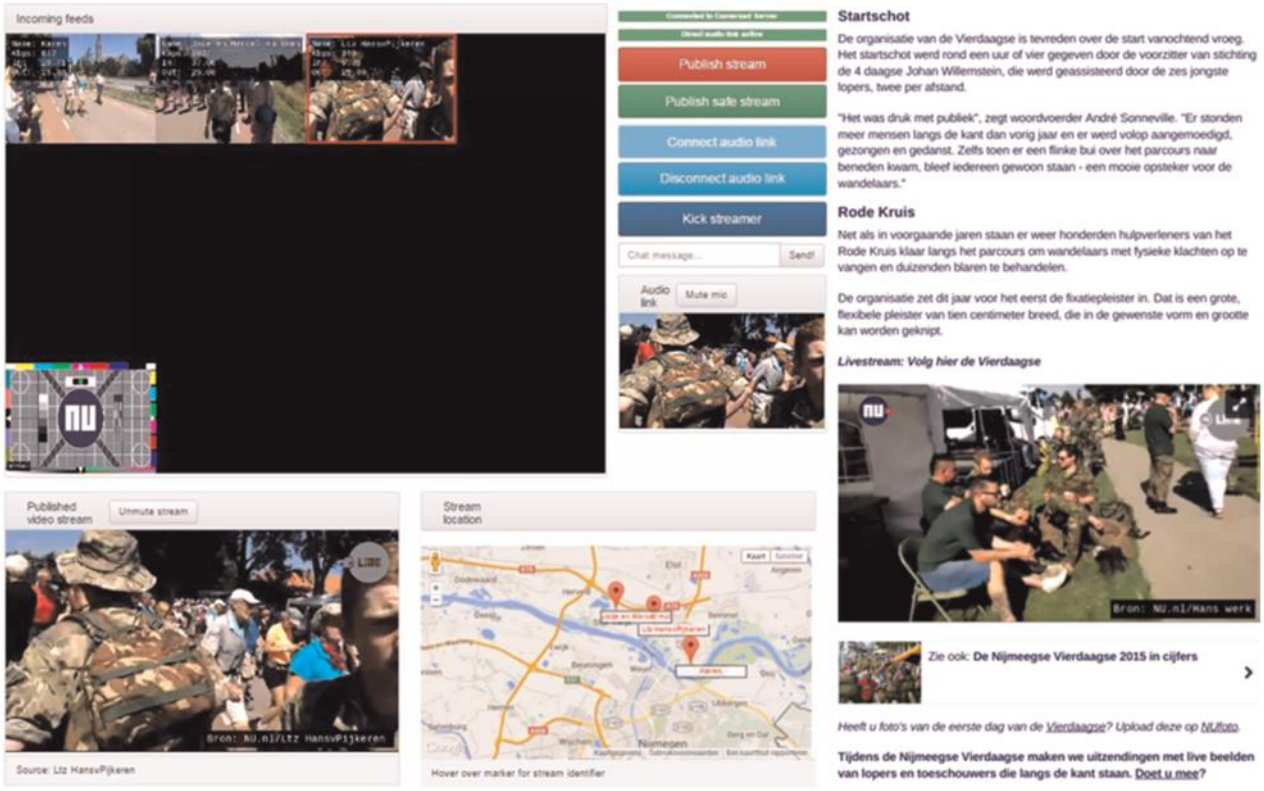


Figure 6 *Second large-scale test during Four Days March in Nijmegen*

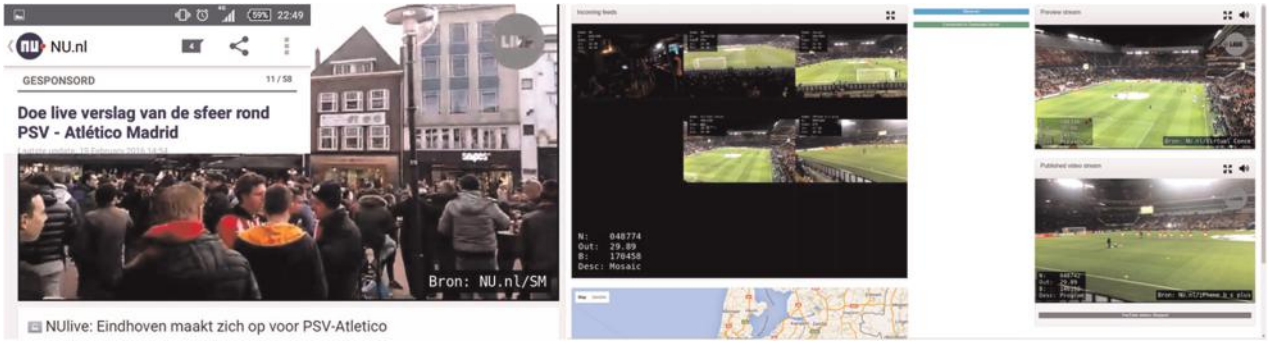


Figure 7 *Final large-scale test during PSV vs Atlético Madrid*

very well covered by professional registration, was limited, and that creating a storyline for the aggregated UGC proved difficult. Still, up to 10,000 viewers viewed the live stream. See Figure 5 for an impression of the dedicated broadcast studio and the incoming streams.

2015 Four Days March

A second large-scale test was performed during the Four Days March. During a period of four days, a dedicated Nu.nl reporter walked among the participants. 50 pre-registered crowdsourcers among the March participants and the audience provided several hours of live streams. Since the March took place close to the Dutch-German border, limited 3G/4G connectivity was experienced at several times. Still, up to 10,000 viewers returned to the live newsfeed during all four days. On the fourth day, the Cameraad system saw a major overhaul, with a new audio engine for: audio mixing and volume control, picture-in-picture feature, stream storage and real-time platform monitoring. Also, a fully automated system deployment approach was incorporated. See Figure 6 for an impression of the incoming streams in the editor web panel.

2016 UEFA Champions League game: PSV vs. Atlético Madrid

A third large-scale test was performed during the UEFA Champions League game between PSV and Atlético Madrid, in February 2016. Close to 60 participants contributed up to 4 hours of video content in 260 individual streams. In addition, two bars in Eindhoven offered continuous live streams to incorporate in newsfeeds. The underlying system saw improvements in the export and storage of streams, with full configuration enabled via the editor web panel. A/V synchronisation of streams was improved to cope with bad connectivity. See Figure 7 for an impression.

Summary and outlook

The development and subsequent testing of the Cameraad CLMS provided us with several valuable insights; first, we noted that great value lies in the ability to cover unscheduled events, leading to a larger number of views. The integration of the Cameraad client library in an already-installed mobile application (in combination with geo-location features), allows a news provider to reach out rapidly to potential crowdsourcers in the neighbourhood of events. The ability to communicate with live stream contributors is highly appreciated by both editors and crowdsourcers, and increases the endurability of the service, i.e. ensuring people return to contribute. In particular for large-scale high-profile events, this communication is a dedicated editorial task. In our subsequent developments, we shall focus on: improving the stream selection process with respect to quality and content, alleviating the process of communication with streamers, annotating stored content for searching, controlling server-side audio settings, and supporting a full H.264 video codec for higher audiovisual quality. We also plan to investigate the application of the Cameraad CLMS in other domains, e.g. video-based surveillance and remote assistance.

Acknowledgments

The research leading to these results has received funding from the Stimuleringsfonds voor de Journalistiek.

References

1 O.Juhlin, O., Engström, A., Reponen, E.: 'Mobile broadcasting: The whats and hows of live video as a social medium'. Proceedings of MobileHCI 2010, 2010, pp. 35–44
2 Kleinberg, S.: 'Live streaming: The next big thing in social media', *Chicago Tribune*, 2015, available at: http://www.chicagotribune.com/lifestyles/ct-so-sociallivestreaming-meerkat-periscope-20150401-column.html, accessed 1 April 2015
3 Hamilton, W., Garretson, O., Kerne, A.: 'Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media'. Proceedings of CHI 2014, 2014, pp. 1315–1324
4 Constine, J.: 'Facebook confirms live broadcasting will soon open to journalists and verified profiles', 12 August 2015, available at: http://techcrunch.com/2015/08/12/facebook-live-livestreaming/, accessed 8 December 2015
5 Pires, K., Simon, G.: 'YouTube live and twitch: a tour of user-generated live streaming systems'. Proceedings of MMSys 2015, 2015, pp. 225–230
6 STEER: 'Exploring the dynamic relationship between social information and networked media through experimentation', available at: http://fp7-steer.eu/experiments/world-cup-rowing/, accessed 7 April 2016
7 IETF RFC 6716: 'Definition of the Opus audio codec'
8 IETF RFC 6386: 'VP8 data format and decoding guide'
9 IETF RFC 5245: 'Interactive connectivity establishment (ICE): a protocol for network address translator (NAT) traversal for offer/answer protocols'

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 48–52

Introduction to *Electronics Letters*

Launched in 1965, just two years before the first IBC, over the last five decades *Electronics Letters* has published around 45,000 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 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

LTE-A compliant multi-band radio and gigabit/s baseband transmission over 50 m of 1 mm core diameter GI-POF for in-home networks

F. Forni[1,2] ✉, Y. Shi[2], H.P.A. van den Boom[1], E. Tangdiongga[1], A.M.J. Koonen[1]

[1]*COBRA Research Institute, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands*
[2]*Technology Department, Genexis, Lodewijkstraat 1A, 5652AC Eindhoven, The Netherlands*
✉ *E-mail: f.forni@tue.nl*

Abstract: The transmission of multiple standard-compliant long-term evolution advance (LTE-A) bands together with a 4-pulse amplitude modulation (PAM) baseband signal over 50-m-long 1 mm core diameter polymethyl methacrylate graded-index plastic optical fibre is demonstrated. Transmission of eight LTE-A 64-QAM bands and a 1.2 Gbit/s 4-PAM baseband signal over the fibre was achieved resulting in an error vector magnitude <8% and pre-forward error correction (FEC) BER $<10^{-3}$, respectively.

Introduction

The deployment of long-term evolution advance (LTE-A) access of the third generation partnership project (3GPP) and its evolution towards 5G presents a number of challenges. Network densification by means of spatial densification (e.g. femto-cell architecture) and spectral aggregation (i.e. intra- and inter-band carrier aggregation) are pursued as solutions. Network densification implies that new in-home networks are needed for an extensive low-cost broadband wired backbone connecting all the femto-cells and supporting the spectral aggregation and the data traffic from the fixed-wired network [1].

Plastic optical fibres (POFs) with their easy 'do-it-yourself' installation capability are an attractive medium to transport 3GPP and other wired and wireless signals simultaneously, as shown in Fig. 1.

Our previous work has shown that a cost-effective solution is the use of multiple LTE bands in parallel with a pulse amplitude modulation (PAM) baseband signal [2], in an in-home scenario transmitted over graded-index (GI) polymethyl methacrylate (PMMA) POF. However, only a limited distance of 20 m link length was reached and strict spectrum allocation is required [3].

Compared with [3], here we report the achievement of a significantly longer distance, hence more suitable for in-home applications, and we increased the LTE-A total throughput using a low-cost laser diode (LD) and a standard receiver. The following signal processing steps are required: first, digital filtering allows us to allocate the baseband and LTE-A spectra with a minimum separation, in order to decrease the mutual interference and to relax the spectrum allocation. Secondly, digital equalisation is applied to the LTE-A bands for optimising the individual LTE-A band performance. Finally, the combination of the higher output power of the LD, optimised at the operating wavelength of 650 nm, and higher receiver sensitivity, allowed the link length to be increased to 50 m.

Experimental setup

The proposed system is based on a simple intensity-modulated direct-detection optical link. The transmission testbed is shown in Fig. 2.

We aimed to maximise the number of transmitted LTE-A bands within the available link bandwidth, in a format-transparent way without any spectrum shifting. Following these criteria, eight widely deployed bands B_i (where $i = 1, \ldots, 8$) with the downlink channel carrier frequency (f_{DW}) allocated between 450 MHz and 1 GHz are studied, as listed in Table 1. To ensure a standard-compliant signal, the LTE-A bands were generated in accordance with the 3GPP defined test model (E-TM) 3.1 using the highest standardised modulation order (i.e. 64-QAM) [4]. Following the E-TM 3.1, the LTE-A signal power is related to the channel bandwidth, hence the equalisation of the individual channels was performed in order to flatten the power spectrum and avoid driving the laser into saturation. The equalisation was carried out employing digital signal processing of each band by allocating the gains G_i (where $i = 1, \ldots, 8$), as listed in Table 1. Thereafter, the equalised LTE-A bands are combined, as shown in Fig. 2*b*.

According to the LTE-A frequency allocation, the frequency range from DC to 450 MHz is unused. In this frequency range a baseband signal is transmitted. Accordingly, a pseudorandom binary sequence (PRBS) 2^7-1 baseband signal is encoded off-line in the 4-PAM format. The symbol sequence is filtered by a 430 MHz digital lowpass filter (LPF), as shown in Fig. 2*b*. Consequently, the LTE-A and baseband signals are generated by two arbitrary waveform generators working as DACs. The DAC output signals are combined and amplified by 10 dB and the resulting signal drives the low-cost LD operating at 650 nm. The LD emitted an optical output power of 5.7 dBm at a wavelength of 650 nm and it was coupled into the fibre by using a ball lens. The optical signal is then transmitted over 50 m of Ø1 mm core PMMA GI-POF, with the fibre loss of 0.2 dB/m at 650 nm [5]. The optical receiver was a Graviton SPD-2 module, consisting of a pin photodiode followed by a transimpedance amplifier. The SPD-2 has an FC connector in which the bare fibre is plugged. The received optical power is equal to −5 dBm, in order to safeguard the p–i–n, and −10.7 dBm for optical back-to-back (oB2B) and 50 m transmission. As shown in Fig. 2*c*, the output signal of the optical receiver was amplified with a gain of 29 dB and acquired by either the baseband 4-PAM or the LTE-A receiver, as shown in the spectrum Fig. 3. For the oB2B test an extra 10 dB attenuator was used before the amplifier, to avoid the amplifier saturation. For the baseband receiver, a digital sampling scope is employed as ADC, a digital LPF with the same bandwidth as in the transmitter, and the 4-PAM receiver, both implemented by MATLAB. For simplicity, no equalisation or precoding was used. The LTE-A signals are received and processed by a vector signal analyser consisting of a spectrum analyser and a Keysight LTE-A software receiver.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 54–56

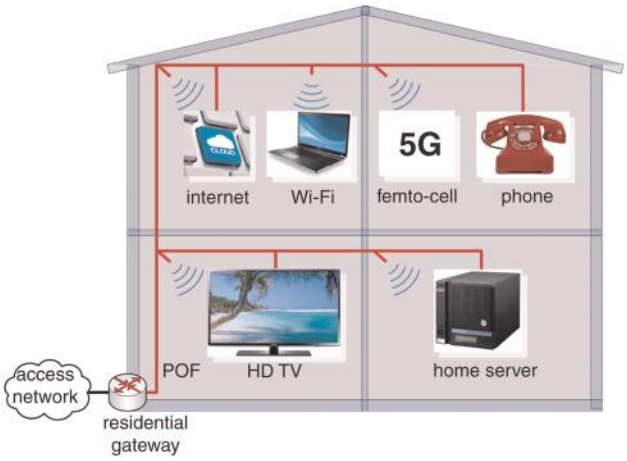


Fig. 1 *POF network for multi-standard wired and wireless fibre in home applications*

Table 1 LTE-A bands and main parameters used in experiments

Bi	31	12	13	14	20	18	19	8
f_{DW} (MHz)	465	738	751	763	806	868	883	944
Bandwidth (MHz)	5		10		20		15	10
Modulation format					64-QAM			
G_i (dB)	−3		0			3	4	2

Experimental results

The LTE-A signal is evaluated in accordance with [6] by taking the error vector magnitude (EVM) value of 8% as threshold. The 4-PAM signal is required to have a pre-FEC BER value of $<10^{-3}$. Initially, the performance of the 4-PAM and LTE-A signals are separately measured as a reference.

Let us consider the 4-PAM performance first. As depicted in Table 2, the 4-PAM transmission over the oB2B is feasible up to a bit rate of 1.54 Gbit/s. When increasing the link to 50 m, the achievable bit rate drops to 1.4 Gbit/s. Moreover, when also the LTE-A signal is co-transmitted, the bit rate decreases further to 1.2 Gbit/s. In case of only 4-PAM transmission, the signal amplitude

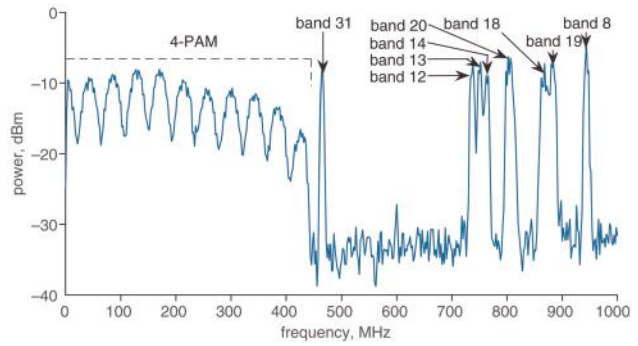


Fig. 3 *Spectrum at receiver amplifier output during simultaneous transmission*

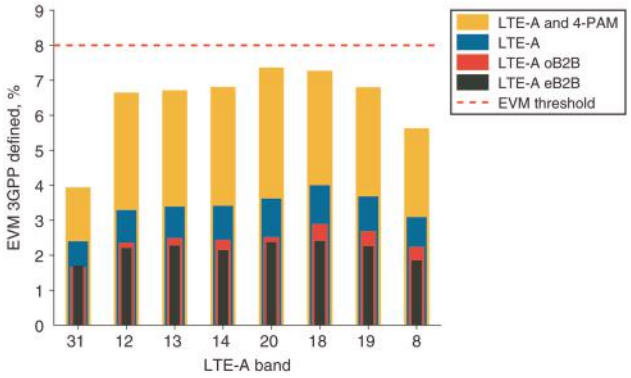


Fig. 4 *LTE-A EVM results, with and without 4-PAM transmission, bands are according to Table 1*

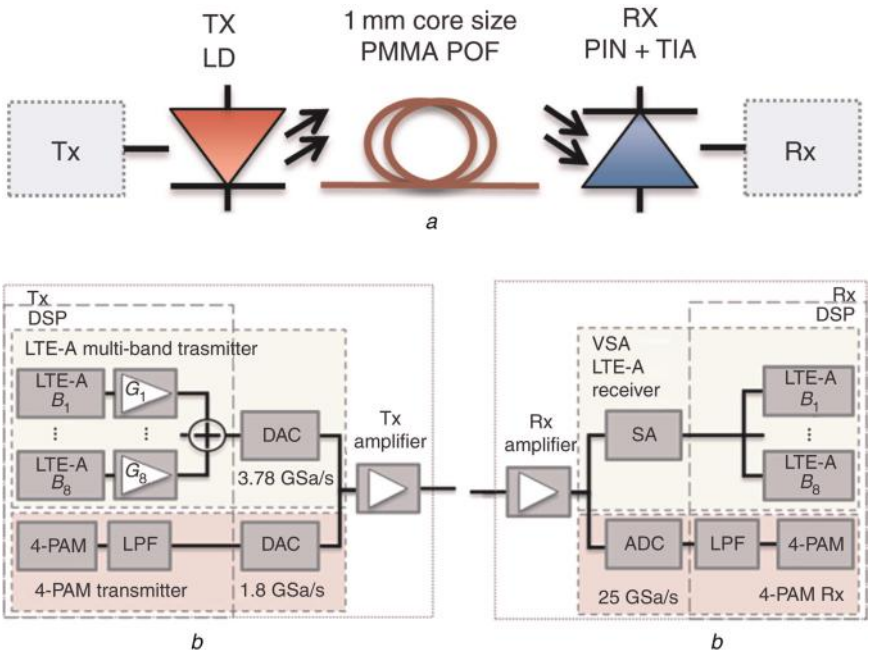


Fig. 2 *Experimental setup diagram consisting of*

a Optical link
b Transmitter
c Receiver

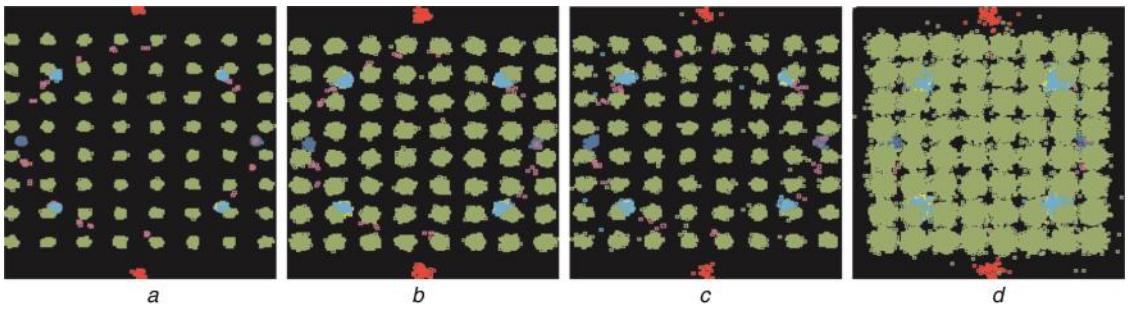


Fig. 5 *Constellation diagram of two received LTE-A bands*

a Band 31
b Band 20
c Band 31 with 4-PAM
d Band 20 with 4-PAM

Table 2 4-PAM transmission experimental results

Link	Bit rate, Gbit/s	BER ($\times 10^{-4}$)
oB2B	1.54	3.6
4-PAM over 50 m	1.4	3.4
4-PAM + LTE-A over 50 m	1.2	8.7

can be adjusted to optimise the throughput. However, when co-transmitting with LTE-As, the 4-PAM amplitude increment can severely affect the LTE-A performance through the non-linearity of the light source. Consequently, during the simultaneous transmission we had to decrease the 4-PAM amplitude, causing the PAM throughput to be lower than the case without the LTE-A bands. Our results show that a minor impairment on the baseband signal bit rate is caused by the LTE-A signals.

Next, the LTE-A multi-band transmission is considered for the electrical back-to-back (eB2B), oB2B, and 50 m link. As depicted in Fig. 4, the eB2B EVM is, in general, below 2.4%. Similar performance among all the LTE-A bands is achieved, thanks to the equalisation technique used. The outermost bands 8 and 31 have the best performance. Moving to oB2B, the EVM increases to values of <2.9%.

Employing 50 m POF makes EVM to increase to <4%. When also the 4-PAM signal is present in the link, the EVMs become <7.4%, which are still better than the 8% allowed EVM. In general, the presence of 4-PAM signal causes an increase in EVM values of LTE-A signals, especially the innermost bands. The lowpass filtering does not necessarily lead to degradation of LTE-A bands spectrally located close to the 4-PAM signals. The nearest-neighbour LTE-A band to the 4-PAM spectrum (i.e. band 31) has the lowest EVM increase, as also shown by the constellation diagram in Fig. 5. Better EVMs can be achieved by decreasing the 4-PAM power, hence limiting interference within the LTE-A bands, albeit the 4-PAM performance decreases. Therefore, for in-home networks, a cost-effective technique for mitigating non-linearity effects on light sources is necessary to further improve the throughput and performance of wired and wireless signals.

Conclusion

In this Letter, we achieved the successful transmission of multiple LTE-A bands and baseband signals over 50 m of thick-core PMMA GI-POF. We have performed the simultaneous transmission of eight LTE-A compliant 64-QAM bands with a total throughput of 478.8 Mbit/s, together with a 1.2 Gbit/s 4-PAM baseband signal. Furthermore, only low-cost components were used without complex digital signal processing and 50 m distance was reached, which could be considered as a good reference length for short distance applications (e.g. single-detached dwelling and apartment). This Letter has demonstrated the feasibility of using POF as in-home network backbone for the LTE-A-based future 5G wireless and wired technologies, which enables a cost-effective approach for supporting the network densification and baseband communications.

Acknowledgment

This research is conducted in the Merging Electronics and Micro and nano-Photonics in Integrated Systems (MEMPHIS) project A2, the Flexible Broadband Communication (FlexCom), supported by the Dutch Technology Foundation STW through the grant 13530.

References

1 Lannoo, B., Dixit, A., Colle, D., *et al.*: 'Radio-over-fibre for ultra-small 5G cells'. 17th Int. Conf. on Transparent Optical Networks, ICTON 2015, Budapest, Hungary, July 2015, pp. 1–4
2 Loquai, S., Kruglov, R., Schmauss, B., *et al.*: 'Comparison of modulation schemes for 10. 7 Gb/s transmission over large-core 1 mm PMMA polymer optical fiber', *J. Lightwave Technol.*, 2013, **31**, pp. 2170–2176
3 Forni, F., van den Boom, H.P.A., Shi, Y., *et al.*: 'Full-service home area networks using plastic optical fibers'. 24th Int. Conf. on Plastic Optical Fibers, Nuremberg, Germany, September 2015, pp. 223–226
4 3-GPP: 'E-UTRA test model 3.1 (E-TM3.1)', LTE – Evolved Universal Terrestrial Radio Access (E-UTRA) Base Station (BS) conformance testing 136.141, 2012, p. 68
5 Ziemann, O., Krauser, J., Zamzow, P.E., *et al.*: 'Optical fibers' in POF handbook, optical short range transmission systems' (Springer-Verlag, Berlin, Germany, 2008, 2nd edn.), p. 92, ch. 2, sec. 2.3
6 3-GPP: 'Error Vector Magnitude', LTE – Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception 136.104, 2012, p. 35

Electrical switching of photoluminescence of single site-controlled InAs quantum dots

A. Schramm[1], E. Koski[2], J.M. Kontio[2] ✉, J. Tommila[1], T.V. Hakkarainen[1], D. Lupo[2], M. Guina[1]

[1]*Optoelectronics Research Centre, Tampere University of Technology, P.O. Box 692, FIN-33101 Tampere, Finland*
[2]*Department of Electronics and Communications Engineering, Tampere University of Technology, P.O. Box 692, FIN-33101 Tampere, Finland*
✉ *E-mail: donald.lupo@tut.fi*

Abstract: Voltage-controlled photoluminescence (PL) switching is demonstrated for single site-controlled InAs quantum dots (QDs) embedded in Schottky–i–n diodes grown by molecular beam epitaxy on nanoimprint lithography patterned GaAs templates. The PL emission was quenched by applying a voltage over the diode structure due to the increased tunnelling rate of charge carriers out of the QDs.

Introduction

In recent years, quantum dots (QD) have found application in novel displays [1], solar cell concentrators [2] and advanced light sources [3]. Furthermore, in all-optical computing, which requires nanosized light sources, QDs are ideal because of their small size, robustness and quantum efficiency. For the development of single QD-based applications, it is crucial to control the location and control the optical properties of the individual QD. With epitaxially grown QDs, lateral positioning can be achieved by patterning the substrates before QD deposition [4]. Control of the emission, for example, switching on and off, can be realised by embedding the QDs in diode structures and using electric fields [5]. In addition, arrays of QDs at controlled distances are a potential building block for QD cellular automata (QCA), in which logical operations are affected by internal reorganisation of charges within the cells of QDs and which enable computation without net charge flow [6]. Voltage control of charge states in QDs after photoexcitation could be a new path to addressing, clocking and reading of QCAs.

In this Letter, we demonstrate voltage-controlled switching of the photoluminescence (PL) of single, site-controlled InAs QDs despite a considerable amount of defects from the regrowth interface close to

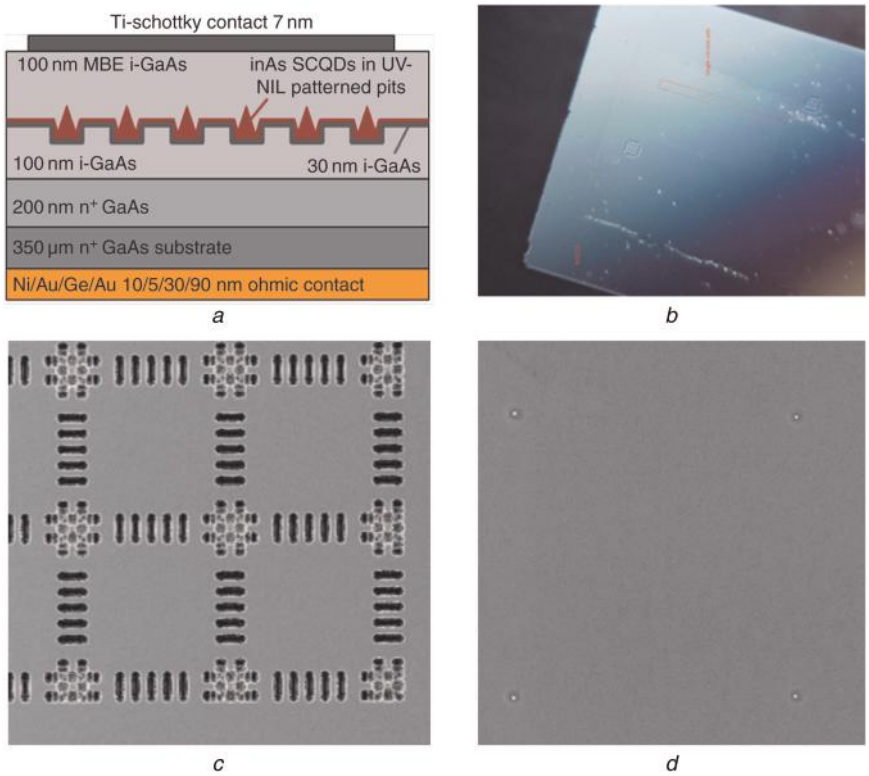


Fig. 1 *Structure of tested devices*

a Layer structure of sample
b Microscope image of processed sample
c SEM picture of marker structure, arrays have size of 30 µm consisting of 5 × 5 QDs
d SEM image of SCQDs having separation of 2.5 µm

the QDs. The electric field is applied parallel to the growth axis. Though the PL emission of site-controlled QDs (SCQDs) can be controlled by the electric field, regrowth interfaces in the junction must also be considered.

Experiments

The site-controlled InAs QDs were grown on a patterned n-GaAs (001) wafer by molecular beam epitaxy (MBE). First, an n-GaAs ($N_D = 10^{18}$ [where N_D is the density of donor atoms]) buffer and an undoped GaAs layer of 200 and 100 nm thicknesses, respectively were grown. Afterwards, a soft ultraviolet nanoimprint lithography process was applied in order to pattern the GaAs surface. Dry etching was used to form pits for QD nucleation as well as alignment marks in the same patterning step. The pits were aligned in a square array with a period of 2.5 μm and diameters of 80, 100 and 120 nm, with a depth of 15–20 nm. In the next step, the samples were chemically cleaned, and the native oxide was removed. Subsequently, the samples were loaded into the MBE chamber. These processes are described in [7]. After patterning, a 30 nm undoped GaAs buffer layer and 1.8 monolayers of InAs were deposited at 470 and 540°C, respectively. For surface characterisation, the sample was unloaded from the MBE chamber after the QD deposition. For micro-PL (μ-PL) studies, the QDs were capped by 100 nm undoped GaAs grown at 540°C. The full layer structure is sketched in Fig. 1*a*. Atomic force microscopy and SEM were used to characterise the surface samples.

To characterise optically SCQDs, the QDs were embedded under a GaAs cap, the sample was loaded into a low-vibration closed-cycle helium cryostat and the μ-PL was measured at 5 K temperature. QDs were optically exited with a laser diode at 640 nm. An approximately 1 μm spot diameter was achieved with a 50× high numerical aperture objective, which was also used to collect the PL signal. A 0.75 m spectrometer with a 1800 lines/mm grating and a cooled Si CCD camera was used to analyse the spectrum.

For electrical measurements, a 7 nm thick semi-transparent titanium Schottky contact on top and an ohmic Ni/Au/Ge/Au contact on bottom were prepared. A microscope image of the processed sample is shown in Fig. 1*b*. The marker structure and the titanium contact are clearly visible. The array markers surrounding the SCQDs (Fig. 1*d*) are depicted in Fig. 1*c*. Each of these areas contains an array of 5 × 5 QDs with a period of 2.5 μm.

Results and discussion

To study the QD PL emission as a function of electric field, we first measured the excitation-power dependent μ-PL of a suitable InAs QD, as shown in Fig. 2. The exciton (X) and biexciton (XX) emission lines are identified according to their PL intensities when excited with different laser powers. The X and XX have slopes of 0.75 and 1.4, respectively, in the log–log depiction of PL intensity against excitation power, which is common for typical InAs QDs [8].

Fig. 3 presents μ-PL measurements of the same single QD against the applied voltage across the Schottky junction. PL quenching is clearly observed when the voltage is made more negative. By −2 V the PL emissions of the X and XX have disappeared. Furthermore, we observe a slight shift of the X PL wavelength against the reverse voltage, as shown in the inset of Fig. 3. The decreasing PL signal is explained by an increased tunnelling rate of charge carriers, mainly electrons due to their lower effective mass, out of the QDs. At a certain point, below −0.8 V in this case, the electrons tunnel faster than they can recombine radiatively. Thus, the exciton peaks are quenched in the PL spectrum. The wavelength shift is attributed to the quantum confined Stark effect, in which the applied electric field modifies the band structure of the QDs resulting in the wavelength shift of the exciton lines. The PL intensities of the exciton and the biexciton against the voltage are shown in Fig. 4. The quenching of the PL signal of the X and XX appears at different voltages (X: −2 V, XX: −1.6 V). This is related to the different saturation

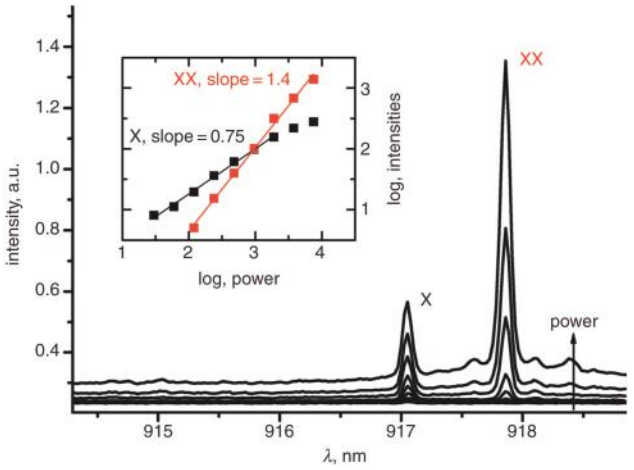


Fig. 2 *Excitation-power dependent μ-PL of InAs QD. Exciton (X) and biexciton (XX) emission lines are identified*

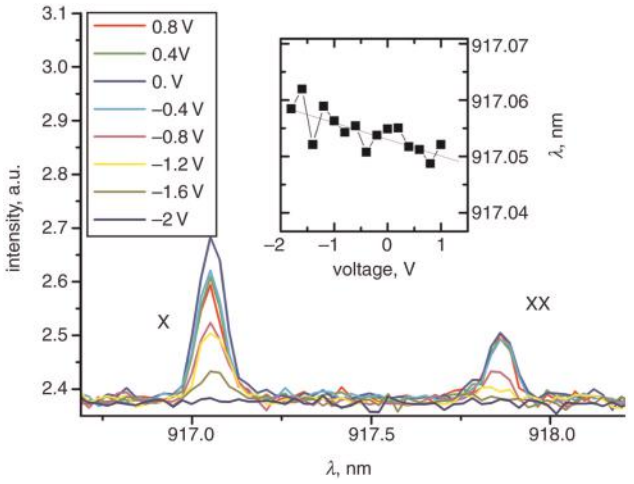


Fig. 3 *μ-PL measurements of single SCQDs against applied voltage. Inset shows wavelength shift of X emission*

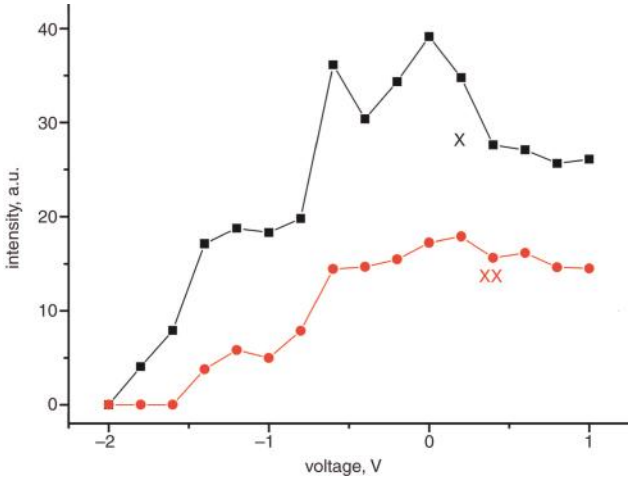


Fig. 4 *Exciton and biexciton PL intensities of single, site-controlled InAs QD against applied voltages*

levels of the exciton and the biexciton. Steps in the PL intensities against voltages are observed for both the exciton and the biexciton. These can be associated to the non-homogeneous

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 57–59

electric field distribution in the junction due to the regrowth interface.

Conclusion

We have demonstrated voltage-controlled PL emission of single, site-controlled InAs QDs embedded in Schottky–i–n diodes. Despite the considerable amount of defects in the regrowth interface as a result of the patterning process, the emission of excitons and biexcitons can be quenched by the applied electric field. This enables the use of voltage control to influence the charge states of QDs after optical excitation. In the future, voltage control of the PL will enable QDs to be used as optical switches and as means to manipulate single-photon sources. In addition, control of the charge states in QD arrays can be a promising path towards QCAs controlled by a combination of electric field and light.

Acknowledgment

We thank the Academy of Finland via the project 'Photonic QCA' (decision number #263594).

References

1 Sun, Q., Wang, Y.A., Li, L.S., *et al.*: 'Bright, multicoloured light-emitting diodes based on quantum dots', *Nat. Photonics*, 2007, **1**, (12), pp. 717–722, doi: 10.1038/nphoton.2007.226
2 Coropceanu, I., Bawendi, M.G.: 'Core/shell quantum dot based luminescent solar concentrators with reduced reabsorption and enhanced efficiency', *Nano Lett.*, 2014, **14**, (7), pp. 4097–4101, doi: 10.1021/nl501627e
3 Tommila, J., Belykh, V.V., Hakkarainen, T.V., *et al.*: 'Cavity-enhanced single photon emission from site-controlled In(Ga)As quantum dots fabricated using nanoimprint lithography', *Appl. Phys. Lett.*, 2014, **104**, (21), p. 213104, doi: 10.1063/1.4879845
4 Schramm, A., Tommila, J., Strelow, C., *et al.*: 'Large array of single, site-controlled InAs quantum dots fabricated by UV-nanoimprint lithography and molecular beam epitaxy', *Nanotechnology*, 2012, **23**, p. 175701, doi: 10.1088/0957-4484/23/17/175701
5 Yakes, M., Yang, L., Bracker, A.S., *et al.*: 'Leveraging crystal anisotropy for deterministic growth of InAs quantum dots with narrow optical linewidths', *Nano Lett.*, 2013, **13**, pp. 4870–4875, doi: 10.1021/nl402744s
6 Snider, G.L., Orlov, A.O., Amlani, I., Zuo, X., Bernstein, G.H., Lent, C.S., Merz, J. L., Porod, W.: 'Quantum-dot cellular automata: Review and recent experiments', *J. Appl. Phys.*, 1999, **85**, (8), pp. 4283–4285, doi: 10.1063/1.370344
7 Tommila, J., Tukiainen, A., Viheriälä, J., Schramm, A., Hakkarainen, T., Aho, A., Guina, M.: 'Nanoimprint lithography patterned GaAs templates for site-controlled InAs quantum dots', *J. Cryst. Growth*, 2011, **323**, pp. 183–186, doi: 10.1016/j.jcrysgro.2010.11.165
8 Kaiser, S., Mensing, T., Worschech, L., Klopf, F., Reithmaier, J.P., Forchel, A.: 'Optical spectroscopy of single InAs/InGaAs quantum dots in a quantum well', *Appl. Phys. Lett.*, 2002, **81**, pp. 4898–4900, doi: 10.1063/1.1529315

Semi-automatic tool for motion annotation on complex video sequences

M.H. Mahmood ✉, *J. Salvi, X. Lladó*

Dept. of Computer Architecture and Technology, Computer Vision and Robotics (VICOROB) Institute, University of Girona, Spain
✉ *E-mail: mhabib.mahmood@udg.edu*

Abstract: Ground truth annotation on motion segmentation (MS) datasets of arbitrary real-life videos is a difficult and challenging task. The research community lacks a standard annotation tool for such datasets, which makes it an open research field. Here an annotation tool is proposed for trajectories in complex videos, which provides a publicly available platform to create and reinforce MS datasets. The user friendly interface allows to refine an initial automatic segmentation result to produce ground truth annotation on all the motions of all the frames of a given sequence. In long videos with multiple rigid/non-rigid motions containing complete occlusion and real distortions, the tool facilitates rapid annotation of motion in a semi-automatic way.

Introduction

Motion segmentation (MS) is an essential building block for many computer vision-based applications. Its most known trajectory-based publicly available datasets are *Hopkins155* [1] and *UdG-MS15* [2]. The trajectory-based *Hopkins155* remains the most extensively utilised MS dataset thus far because of the inherent sparse trajectories present in it and the availability of a standard evaluation metric, which made it easy for algorithms to be compared at an equal scale. However, the limitations in *Hopkins155* became apparent when algorithms started presenting results with <1% of misclassification. It happened because *Hopkins155* contained short videos, i.e. 30–40 frames, of mostly synthetic sequences with no missing data, whereas real world scenes contain complex motions over hundreds of frames with multiple occlusions and other real-life noise. These limitations in *Hopkins155* also arise from the fact that ground truth annotation of sparse feature trajectories in long sequences with multiple complex motions, with the presence of real distortions and missing data, is a challenging task.

In a trajectory-based dataset, given a video sequence, the moving objects are tracked through sparse trajectories to build a trajectory matrix $\boldsymbol{W}_{2f\times p}$, where f is the number of video frames and p is the number of tracked feature points (trajectories). The sparsity of the tracked feature points on the moving objects varies with respect to the desired density of coverage. Independent of the density of coverage, a trajectory-based dataset with long natural sequences can only be created in the presence of a trajectory annotation platform. Though region-based labelling platforms have recently been proposed [3, 4], the community lacks a standardised trajectory annotation platform, where sparse trajectories on long sequences of natural scenes can be annotated.

In this Letter, we present an easy-to-use semi-automatic feature trajectory annotation tool (TAT) for efficient labelling of complex motions in long video sequences. Our tool works independently of the sparsity of trajectory density coverage. Therefore, it can provide a common platform for researchers, where any tracking result could be used to build trajectory-based MS datasets. We used here the state-of-the-art large displacement optical flow (LDOF) tracker [5] for this purpose. TAT also gives the user the flexibility to employ a semi-automatic or a manual mode, where an initial motion label is assigned to all trajectories. To do this, we used the recently proposed OB algorithm [6], which allows to incorporate an initial segmentation result. A comprehensive evaluation of the TAT tool when creating the ground truth of a state-of-the-art MS dataset [2] is presented, which exhibits that any motion spanning even over 100 or more frames might be labelled in <7.5 min on average.

TAT overview

The scheme of the TAT annotation tool is presented in Fig. 1. It is developed using MATLAB libraries for graphical user interface (GUI), with all the resources publicly available on-line at http://www.udgms.udg.edu/TAT/. The tool offers two modes; *semi-automatic* or *manual*. In *semi-automatic* mode, the OB [6] MS algorithm is used as label initialisation, and then the annotation is refined to form a final ground truth. It should be noted that TAT provides the flexibility to use any MS algorithm for initialisation. In *manual* mode, the label initialisation is done by annotating all trajectories as background by default.

TAT supports different annotation options. *Point-wise:* point by point labelling; *trajectory-wise:* trajectory labelling by selecting a single point on it; and *region of interest (ROI)-wise:* by labelling in one shot a cluster of trajectories in an ROI. The inputs are the *video frames* to be labelled along with the *trajectory matrix* $\boldsymbol{W}_{2f\times p}$, which is the output of the LDOF [5] tracker algorithm. An input *label vector* $\boldsymbol{L}_{p\times 1}$ is also used, which changes according to the selected initialisation mode. When using the semi-automatic mode, the output of the OB MS algorithm is used as a label vector, while in manual mode, all the trajectories are initialised by default with the background label. These input variables are converted into three indexes: label, trajectory and frame. Each index is dependent on the annotation update.

The label index uses the label vector $\boldsymbol{L}_{p\times 1}$ to keep track of the ID tagged with each motion independent of the initialisation mode. When a trajectory on a new moving object is selected and annotated, a new label is added in the label index. As more trajectories on a motion are annotated, they form a trajectory cluster of the object across all the frames. If at any point during annotation, an already existing label ID is given to a new trajectory cluster then the two clusters are merged. This provision is especially useful if, mistakenly, two trajectory clusters with unique IDs are formed of the same moving object then both can be merged in one go. The trajectory index is formed by using the trajectory matrix with the label index. It keeps a log of association of each trajectory with its annotated label index ID. This is specifically useful in the display engine as all the trajectories belonging to a unique label index ID are displayed in a unique

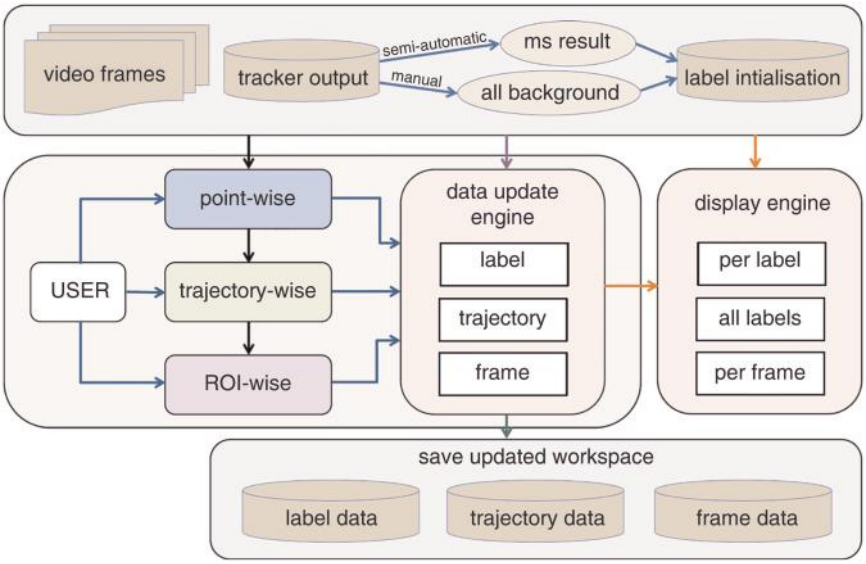


Fig. 1 *TAT flow-diagram, input block at top with label initialisation modes, data update and display engines in middle with save option at bottom*

colour. The trajectory index is essentially used to form the annotation result, *Updated Label vector* $\boldsymbol{Lu}_{p\times 1}$. The frame index is formed by the input frames of the video sequence along with the trajectory index and the label index. This index is the back bone of the display engine, which facilitates swift annotation in TAT while providing a visual check on the correctness of each annotated motion at the same time. A log of each trajectory with reference to its updated label present in each frame is kept in this index. The index contains multiple trajectory clusters with associated labels per frame. While scrolling through the frames this index shows each label ID with reference to the selections made in the display engine modalities. In TAT, these indexes provide the structure for data update and display engines to be efficiently used. The display window of the GUI is shown in Fig. 2, where a frame with trajectory overlay, after complete annotation on a few moving objects, can be seen.

The interface is kept simple with each annotation and display modality directly available for the user. In this way, the tool shares enhanced control with the user so as to maintain flexibility. Fig. 3 presents the application of the semi-automatic mode of TAT in a traffic sequence, using OB [6] for label vector initialisation. Depending on the performance of the OB segmentation, the initialised $\boldsymbol{L}_{p\times 1}$ will have some motions, partially or completely, correctly labelled. Although, by using the semi-automatic mode, the overall user annotation time is reduced, the time needed to refine the remaining trajectories is dependent on OB's failures. To reduce the refinement time needed to correct OB failures, the trajectory-wise and ROI wise options are preferable as they select complete clusters in one go. The choice of a trajectory selection option is dependent upon the type (rigid or non-rigid), size and shape of the moving object, its distance from the camera and its position in the field of view. As a thumb rule, the ROI selection option should be preferred in the homogeneous regions of the moving objects. The use of the ROI-wise modality might induce an error in non-rigid, small or irregularly shaped objects, as, neighbouring trajectories that belong to either background or other moving objects, can mistakenly get selected. Therefore, for all such objects, trajectory-wise option should be used. With reference

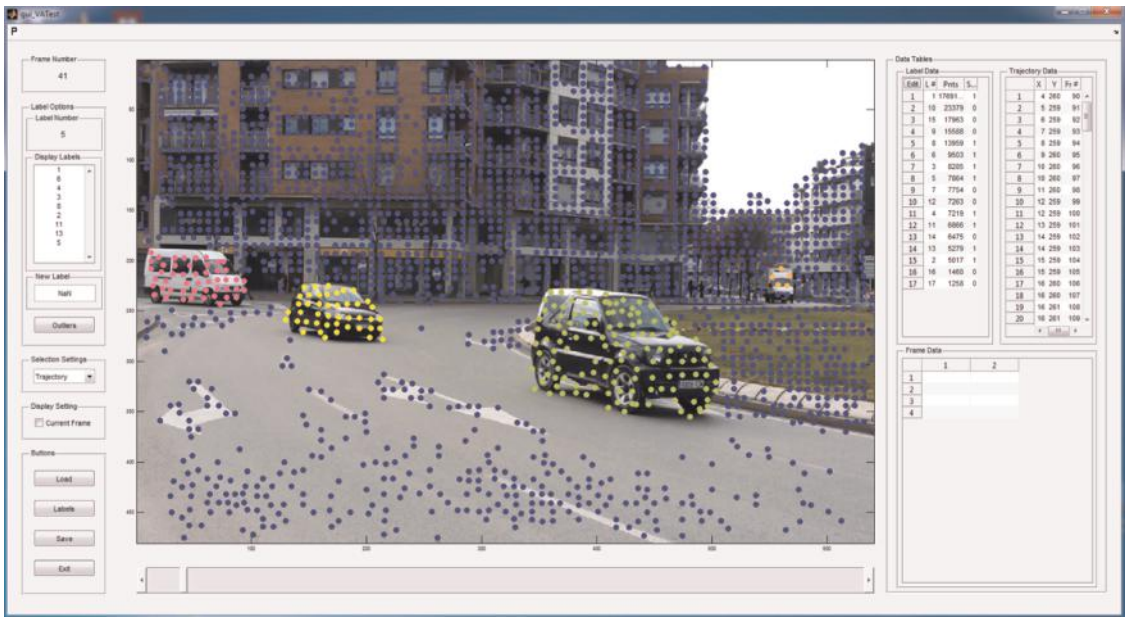


Fig. 2 *TAT GUI, with display window in middle, labelling modes and selections on left and selected indexes on right*

Fig. 3 *Sample frames of video sequence in Traffic group after complete annotation*

a Frame 90
b Frame 160
c Frame 240
d Frame 390
e Frame 510
f Frame 600

to Figs. 3*b* and *e*, when the moving objects are about to be occluded or are about to enter the field of view, the selection of trajectories on the border of each motion region becomes critical.

Evaluation

We evaluate the performance of the TAT tool when creating the ground truth of the MS dataset *UdG-MS19*. This novel database, whose seminal work was presented as *UdG-MS15* [2], contains 19 long video sequences of natural scenes with multiple motions of different types spanning over hundreds of frames having partial and complete occlusions. As compared with Hopkins, these new challenges presented in this dataset provide a new benchmark for the community.

To quantify the performance of the tool, the total user annotation time, UT is used. This is the cumulative time required to annotate each motion in a sequence. The time taken to completely label each motion is dependent upon the motion type, camera motion, moving object size and frame length of each trajectory. All these traits were extensively tested and the obtained evaluation results are shown in Table 1. In order to effectively analyse TAT, the sequences were grouped into four distinct motion types; *traffic* (MT1), *people-traffic* (MT2), *relative camera motion-traffic* (MT3) and *people-traffic-camera jitter motion* (MT4). The given names exhibit the motion types of each group. We used the *F*-score [6] to quantify the correctly classified motions by OB initialisation. This score takes *sensitivity* and *precision* into account. Its scale is in percentage with a maximum of 100%, which would mean that all motion trajectories in a sequence were correctly segmented. To acquire a deeper insight into the annotation time, besides UT, two more time measures were used: UT_s, average time per sequence and UT_m, average time per motion. All the subscripts in our Letter denote averages, represented as s*:* per sequence; t*:* per trajectory; and m*:* per motion.

Table 1 TAT evaluation results. Acronyms are MT: motion types, *S*: number of sequences, *M*: number of motions, *F*: number of frames, *T*: number of trajectories, OB: OB *F*-score, UT: total user time, All subscripts are averages; s: per sequence, t: per trajectory, m: per motion

MT	*S*	*M*	M_s	F_t	T_m	OB	UT	UT_s	UT_m
MT1	5	55	11.0	44.5	179.2	62.3	435	87.0	7.9
MT2	7	63	9.0	100.9	126.6	50.8	290	41.4	4.6
MT3	4	34	8.5	42.6	218.9	13.0	275	68.8	8.1
MT4	3	49	16.3	97.6	82.7	38.3	445	148.3	9.1

Observing Table 1, one can see that it took less UT_s and UT_m to annotate the *MT2* group of sequences. Even though the average frame length of the trajectories in *MT2* were long, this group took less time as it had less motions per sequence, less trajectories per motion and got more than 50% of motions annotated correctly with the OB initialisation. UT_s and UT_m were high for the *MT4* group as the motions per sequence were doubled and OB initialisation failed on more than 60% motions. Here, less number of trajectories per motion also indicated that the size of objects in the *MT4* group was small, which made the time efficient ROI option unusable. The groups *MT1* and *MT3* took almost the same amount of time on average but the reasons they did so were quite different. In *MT1*, with 62% of motions correctly annotated by the OB initialisation, UT should have decreased, but due to more trajectories per motion and more motions per sequence, the overall annotation time increased. In contrast, *MT3* had less motions per sequence so UT should have decreased, however, due to small trajectory frame length and bad OB initialisation the overall annotation time increased for these video sequences. The overall average UT was 7.2 min in a total of 19 sequences with over 800 frames and 10 motions, per sequence on average.

This evaluation gives an insight about the usage of the annotation tool. It is apparent that the semi-automatic modality speeds up the

annotation process. The speed up factor depends on $\boldsymbol{L}_{p\times 1}$ initialisation, a better initialised label vector results in less annotation time. The ROI option is useful if large non-rigid objects with less occlusion are present in the sequence, as they have motion regions with higher area and density coverage. In small, non-rigid or region borders of objects, it is better to use the trajectory-wise selection option. Though this option is not as fast as ROI in terms of user annotation time, the precision it brings is essential for accurate labelling of these difficult motions.

Conclusion

The creation of ground truth in trajectory-based MS datasets is a challenging task, especially in the presence of long real-life sequences with multiple motion types and large frame length. In this Letter, we presented our semi-automatic TAT for complex videos. It enables the community to create the ground truth of an MS dataset on a standardised publicly available platform. We demonstrated that the modalities kept in our tool are flexible, hence the use of any tracker output, and an initialisation from any state-of-the-art MS algorithm, is supported. We also provided an evaluation of our tool when it was used to create the annotations on a novel dataset *UdG-MS19*. The evaluation results showed that the platform can produce rapid annotations on long videos with minimal time requirements, which can benefit the MS research community.

Acknowledgments

This work was supported by the FP7-ICT-2011-7 project PANDORA (Ref 288273), RAIMON (Ref CTM2011-29691-C02-02) and NICOLE (Ref TIN2014-55710-R). M. H. Mahmood is supported by an FI grant associated with RAIMON project.

References

1 Tron, R., Vidal, R.: 'A benchmark for the comparison of 3-D motion segmentation algorithms'. Proc. CVPR, Minneapolis, Minnesota, USA, June 2007, pp. 1–8
2 Mahmood, M.H., Zappella, L., Díez, Y., Salvi, J., Lladó, X.: 'A new trajectory based motion segmentation benchmark dataset (UdG-MS15)', *LNCS-PRIA*, 2015, **9117**, pp. 463–470
3 Bianco, S., Gianluigi, C., Paolo, N., Raimondo, S.: 'An interactive tool for manual, semi-automatic and automatic video annotation', *Trans. CVIU*, 2015, **131**, pp. 88–99
4 Vondrick, C., Patterson, D., Ramanan, D.: 'Efficiently scaling up crowd sourced video annotation', *IJCV*, 2013, **101**, (1), pp. 184–204
5 Brox, T., Malik, J.: 'Large displacement optical flow: descriptor matching in variational motion estimation', *Trans. PAMI*, 2011, **33**, (3), pp. 500–513
6 Ochs, P., Malik, J., Brox, T.: 'Segmentation of moving objects by long term video analysis', *Trans. PAMI*, 2014, **36**, pp. 1187–1200

Fully direct write dispenser printed sound emitting smart fabrics

Y. Li ✉, *R. Torah*, *K. Yang*, *Y. Wei*, *J. Tudor*

Department of Electronics and Computer Science, Faculty of Physical Sciences and Engineering, University of Southampton, Southampton, SO17 1BJ, United Kingdom
✉ *E-mail: Yi.Li@Soton.ac.uk*

Abstract: Direct write dispenser printed sound-emitting smart fabrics are reported for creative applications and wearable electronics. Using dispenser printing, sound emission can be easily added to various fabric substrates, from industrial coated architectural fabrics to everyday woven polyester cotton fabrics. Two different types of sound emission have been realised on fabrics: a piezoelectric buzzer and an electromagnetic planar spiral speaker with an external magnet. These two planar sound emitting devices are flexible and simpler than conventional three-dimensional electromagnetic and electrostatic sound emitting devices. Furthermore, the spiral-based planar electromagnetic speakers allow interactive applications by changing the distance between the magnet and the speaker. The theory and the manufacturing technology of the direct write printed fabric buzzer and speaker are reported.

Introduction

In the past five years, printed smart fabrics research, especially by screen printing, has been growing significantly and has achieved several major milestones, such as low temperature (<150°C) processable functional materials compatible with fabrics. Functions achieved are: conductive [1], dielectric [2], piezoelectric [3], electroluminescent [4], sacrificial [5] and strain sensing [6]. In addition, the challenge of printing functional layers of about 20 μm thickness on polyester cotton fabrics, which has a surface roughness of the order of 150 μm, is met by printing a suitably thick primer interface layer on the fabric before additional layers are printed. A screen printable interface layer to support the subsequent functional layers has overcome the barrier of high roughness of woven fabrics [7]. However, sound emission achieved by printing on fabric has not been reported previously.

The fabrication technique used in this current Letter is direct write pneumatic dispenser printing. The fabrication flow diagram, shown in Fig. 1, represents the steps to direct write dispenser print multiple functional layers on a woven fabric substrate; three different functional materials are used in this example as shown in Fig. 1. Dispenser printing offers the benefit that a screen is not needed and the layers are directly printed from the PC design. Moreover, it accepts a much wider rheological functional inks/paste range compared with the major alternative fabrication techniques of screen or inkjet printing. Dispenser printing also allows variable print resolution by changing the nozzle size, and has the same design flexibility as inkjet printing. Lower gauge nozzles with a higher dispensing pressure can obtain a wider/thicker printed line/layer with a single deposition pass. Dispenser printer fabrication of electronic devices on non-fabric substrates was comprehensively studied in [8]. Dispenser printed electronic devices have been reported [9, 10]; however, dispenser printing has been previously applied to smart fabrics, nor printed sound emitting devices. This Letter, in particular, targets the vision of future interactive fabrics for the creative industries, where interactive fabric products are designed and dispenser printed locally in the studio. This novel direct write dispenser printed smart fabrics system impacts the creative industries facilitating complete freedom of customisation and the smart fabrics integration with any designed patterns on standard commercially available fabrics. This Letter, reports for the first time the realisation of sound emission on fabrics fabricated by dispenser printing.

Theory and design of the printed fabric sound emitting devices

There are two types of sound emitting devices which can be printed as planar structures on fabric: (i) a piezoelectric buzzer comprising three functional layers in a sandwich structure of a piezoelectric layer in between two conductive electrode layers as shown in Figs. 2*a* and *b*. Figs. 2*a* and *b* show piezoelectric layer deformation when excited electrically is used to create sound. It is limited in frequency response and can therefore only be used as a tweeter-like device. The sound volume performance depends on the d_{33} value of the piezoelectric layer and the total area of the device. Any pattern of a piezoelectric buzzer can be designed providing the device sandwich structure is maintained. (ii) A planar spiral speaker consisting of a conductive coil attached to a membrane and a non-contact magnet, as shown in Figs. 2*c* and *d*. The sound emission is generated by the membrane vibration. The conductive coil can be implemented as a planar spiral coil which is equivalent to a wired coil and the fabric substrate acts as the vibrational membrane as shown in Fig. 2*e*. The design can be any closed-loop conductor pattern.

Method and materials for the printed sound emitting devices on fabric

The direct write dispenser printer used in this Letter was developed at the University of Southampton with customised three-axis linear stages and a Musashi ML-808FXcom dispenser controller. The dispenser operates using a continuous extrusion mode to deposit a series of adjacent lines, which then coalesce due to capillary effects to create a uniform film. Three different fabric substrates are selected in this Letter providing increasingly difficult surfaces to print on.

Fabric: Fabric 1, FRONTLIT II FR, is a pre-coated smooth surface architectural fabric, supplied by Mehler (http://www.mehler-texnologies.com); fabric 2, Back Lighttex FR, is a woven architecture fabric, supplied by Berger (http://www.bergertextil.com); and fabric 3 is a standard 65/35 polyester cotton fabric, supplied by Klopman International (www.klopman.com). Fabrics 1 and 2 are mainly used in the architectural decorative industries and fabric 3 is the most widely used fabric for clothing.

The Best of IET and IBC 2016–2017, 2016, Vol. 8, pp. 64–66

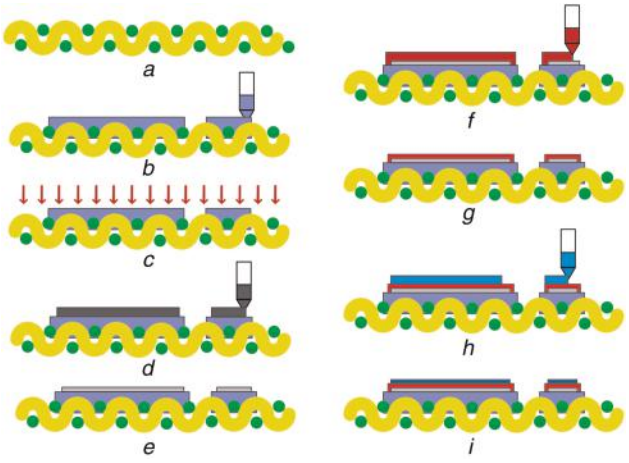


Fig. 1 *Cross-sectional view: fabrication flow diagrams of direct write dispenser printed three different functional layers on woven fabrics*

a Cross-sectional view of untreated woven fabric
b Dispenser printing through nozzle to deposit interface layer on untreated woven fabric substrate
c UV curing to solidify printed interface layer
d Dispenser printing through nozzle to deposit first layer on top
e Thermal curing to solidify first functional layer
f Dispenser printing through nozzle to deposit second layer on top
g Thermal curing to solidify second functional layer
h Dispenser printing through nozzle to deposit third layer on top
i Thermal curing to solidify third functional layer on top to complete three-layer device fabrication process

Pre-treatment: An ultraviolet (UV)-curable interface ink, Fabink-IF-UV4 (Smart Fabric Inks, http://www.fabinks.com), is printed on fabric 3 following a pre-defined pattern to reduce surface roughness (~150 μm) and present a smooth surface for the subsequently printed layers. The interface is only printed where the subsequent conductive layer is needed to maximise the fabric's flexibility and breathability. However, fabric 1 already benefits from its pre-coating and does not require a primer layer as it is sufficiently smooth. Fabric 2 also does not need an interface layer due to the 100% polyester yarn content which helps to prevent the printed functional material bleeding before thermal curing.

Functional materials: The following functional pastes were used and supplied by Smart Fabric Inks: (i) silver paste, Fabink-TC-AG1, was chosen for the conductive electrodes in the lead zirconate titanate (PZT) buzzer and spiral coil in the electromagnetic speaker; (ii) PZT paste, Fabink-PolyPZT, was chosen for the PZT layer in the PZT buzzer; and (iii) the dielectric paste, Fabink-TC-D1, was chosen for the metal–insulator–metal cross-over in the spiral speaker.

Printing of the sound emitting devices

The printing parameters and curing conditions are described in this Section. An 18-gauge and 840 μm inner diameter tapered nozzle from Fisnar (http://www.fisnar.com) was used to dispense the interface layer with a dispensing pressure of 30 kPa and a vacuum level of 0.7 kPa. Using a tapered nozzle allows an increased fluid volumetric flow rate compared with the straight untapered dispensing tips. The vacuum level is used to prevent dripping when not dispensing. The gap between the printing nozzle and the substrate is ~100 μm. The stages moved at 15 mm/s. The same dispensing gap and stage movement speed were used for each printed layer. A single printed layer is sufficient to obtain a smooth interface surface which is cured with an UV dose of 1500 mJ/cm^2 in an UV chamber. The silver layer was dispensed using a 25-gauge and 250 μm inner diameter tapered nozzle with a dispensing pressure of 25 kPa and a vacuum level of 0.7 kPa. A higher gauge nozzle was used compared with the interface paste because the silver paste has a lower viscosity. The silver layer was thermally cured for 10 min at 130°C in a box oven. The piezoelectric paste is based on PZT and was dispensed with the same nozzle as the silver paste but with a dispensing pressure of 160 kPa and vacuum level of 1 kPa. The printed PZT layer was thermally cured for 20 min at 90°C in a box oven. Finally, the dielectric layer was dispensed with the same tapered nozzle and curing conditions as the silver paste but with a dispensing pressure of 40 kPa and vacuum level of 1 kPa. A poling process is required to activate the piezoelectric properties of the printed PZT layer. The poling conditions were 4.5 MV/m at 150°C for 10 min. The

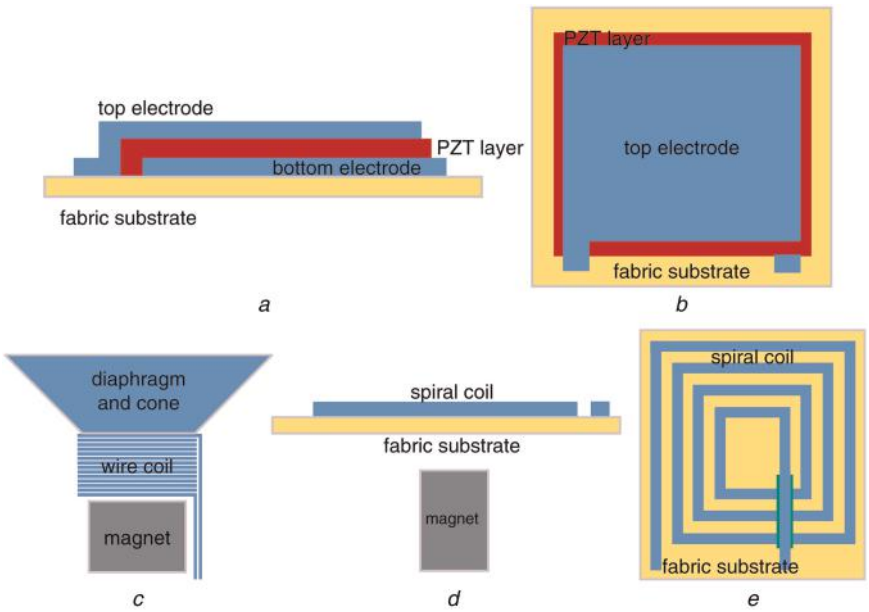


Fig. 2 *Cross-sectional and plan view of schematic diagrams*

a Cross-sectional view of piezoelectric buzzer on fabric substrate
b Plan view of piezoelectric buzzer on fabric substrate
c Cross-sectional view of conventional electromagnetic speaker
d Cross-sectional view of planar spiral speaker based on electromagnetism on fabric substrate
e Plan view of spiral speaker design based on electromagnetism

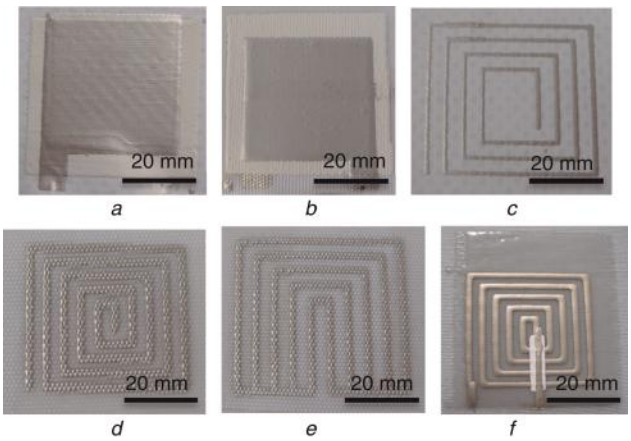


Fig. 3 *Plan view of dispenser printed fabric PZT buzzers and spiral speakers*

a Buzzer on fabric 1 with printed interface layer
b Buzzer on fabric 2 with printed interface layer
c Speaker on fabric 1 with printed interface layer
d Speaker on fabric 2 with printed interface layer
e Speaker on fabric 2 with printed interface layer
f Speaker on fabric 3 with printed interface layer

d_{33} value was measured to be 46 pC/N using a Take-Control PM35 piezometer.

Demonstration of the printed fabric sound emitting devices

Both the dispenser printed PZT buzzer and the spiral speaker are realised on three different fabric substrates. Examples are shown in Fig. 3. The PZT buzzer was tested with a 40 MHz arbitrary waveform generator (TGA1241, TTi Ltd.) connected to a power amplifier (Labworks Inc.). The test showed that the ambient background noise (~40 dB) was higher than the sound level produced by the PZT buzzer. The spiral speaker was tested by connecting an audio source (MP3 player) through a 20 W audio amplifier (PULSE-Audio Ltd SDA20). In addition, a magnet is needed to vary the sound volume by changing the local magnetic flux. A notable sound pressure level, ~50 dB, can be produced in a standard office environment. However, only a high-pitched sound source generates a clear output, such as the '*Four Seasons*' by Vivaldi, since the planar spiral speaker lacks a cone to enhance the output of low-pitch sound waves, such as the human voice.

Conclusion

This Letter introduces a novel fabrication technique of dispenser printing to create sound emitting smart fabric devices. Two different types of dispenser printed sound emitting devices on three different fabric substrates were used to demonstrate the dispenser printing technique. The printed devices demonstrate the advantages of direct patterning the functional devices on the fabrics following a computer-defined pattern without needing a pre-made mask. This printing system can be used to realise interactive smart fabrics locally in the creative workshop with a higher degree of personalisation and in a rapid prototyping manner.

Acknowledgment

This work was supported by the EU under the FP7 project CREATIF, grant number CP-FP-INFSO-FP7-610414.

References

1 Locher, I., Troster, G.: 'Screen printed textile transmission lines', *Tex. Res. J.*, 2007, **77**, (11), pp. 837–842
2 Neral, B., Turk, S.S., Voncina, B.: 'Properties of UV-cured pigment prints on textile fabric', *Dyes Pigments*, 2006, **68**, (2–3), pp. 143–150
3 Almusallam, A., Torah, R.N., Zhu, D., Tudor, M.J., Beeby, S.P.: 'Screen printed piezoelectric shoe-insole energy harvester using an improved flexible PZT-polymer composites'. IOP Publishing, Journal of Physics: Conf. Series 476, PowerMEMS 2013, London, UK, December 2013, p. 012108
4 Sloma, M., Fanczak, D., Wroblewski, G., Mlozniak, A., Fakubowska, M.: 'Electroluminescent structures printed on paper and textile elastic substrates', *Circuit World*, 2014, **40**, (1), pp. 13–16
5 Wei, Y., Torah, R., Yang, K., Beeby, S., Tudor, J.: 'A screen printable sacrificial fabrication process to realise a contilever on fabric using a piezoelectric layer to detect motion for wearable applications', *Sens. Actuators*, 2013, **203**, pp. 241–248
6 Perc, B., Kuscer, D., Holc, J., Belavic, D., Svetec, D.G., Guermonprez, P., Kosec, M.: 'The processing and characterisation of a strain sensor on a textile'. 41st Int. Symp. on Novelties in Textiles, Ljubljana, Slovenia, May 2010
7 Whittow, W., Li, Y., Torah, R., Yang, K., Beeby, S., Tudor, J.: 'Printed frequency selective surface on textiles', *Electron. Lett.*, 2014, **50**, (13), pp. 916–917
8 Ho, C.C.: 'Dispenser printed zinc microbattery with an ionic liquid gel electrolyte'. PhD thesis, Materials Science & Engineering, UC Berkeley 2011
9 Ho, C.C., Evans, J.W., Wright, P.K.: 'Direct write dispenser printing of a zinc microbattery with an ionic liquid gel electrolyte', *J. Micromech. Microeng.*, 2010, **20**, p. 104009
10 Madan, D., Chen, A., Wright, P.K., Evans, J.W.: 'Dispenser printed composite thermoelectric thick films for thermoelectric generator applications', *J. Appl. Phys.*, 2011, **108**, p. 034904

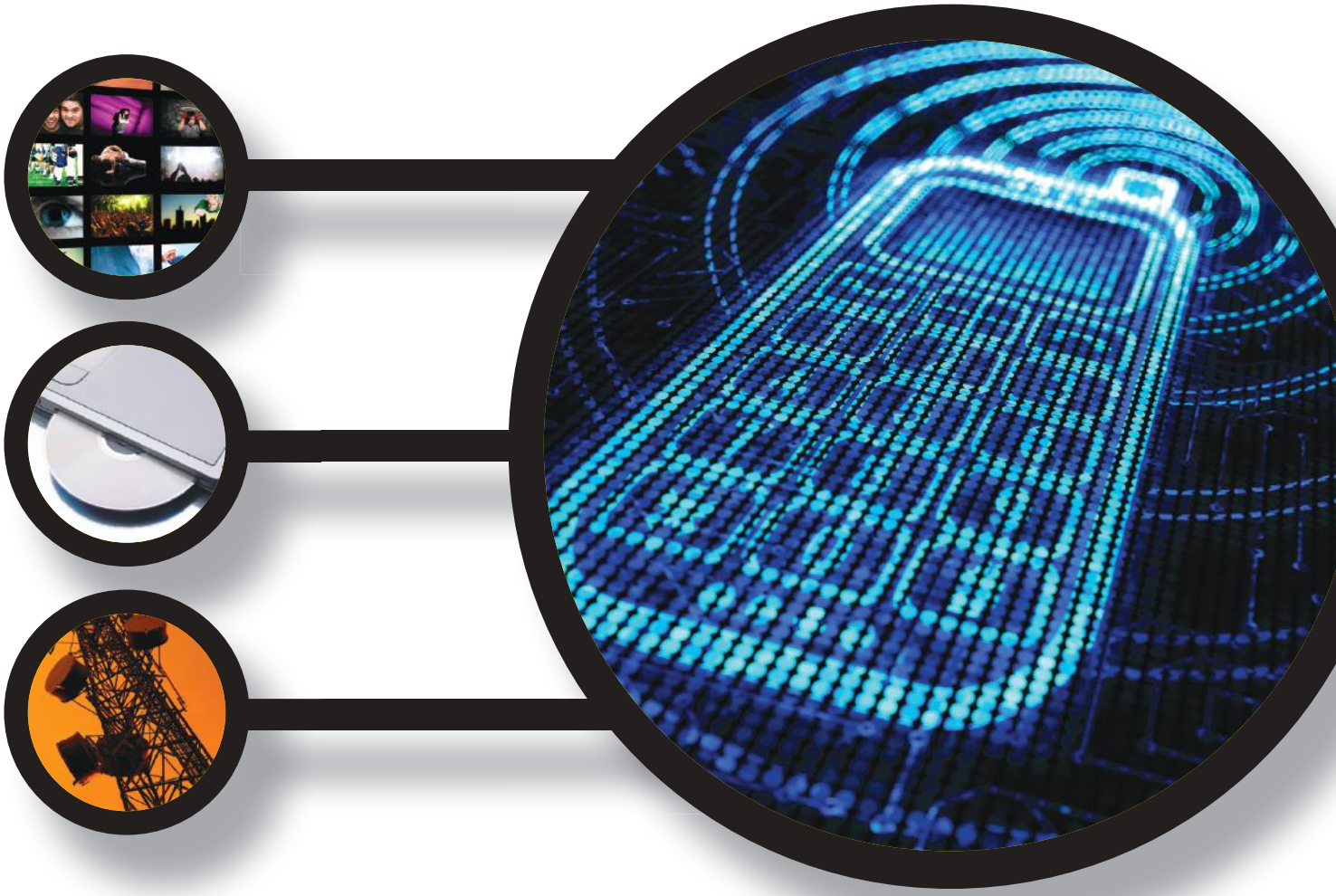
Multimedia Communications Network

IET Communities
The Institution of Engineering and Technology

The IET **Multimedia Communications Network** is the go to place for the multimedia communications community to keep up to date with latest developments and insights as well as connecting with other professionals.

This Technical Network covers the areas of creating, distributing and consuming dynamic media for multiple devices, in multiple locations. Connect with the IET **Multimedia Communications Network** **online**, at **events** and through **content** to make sure you don't miss out.

Be a part of your engineering community today!



www.**theiet**.org/multimedia

IET Sectors
The Institution of Engineering and Technology

Information & Communications

YOUR CHANCE TO GET INVOLVED

The IET Information & Communications Sector is working collaboratively with Industry, Academia and Government to engineer solutions for our greatest societal challenges.

- **Connected Data**
- **Cyber Security**
- **Digital Innovation**
- **Ubiquitous Computing**

If you have a project idea or activity you would like to explore with the support of the IET please get in touch.

Email: **sectors@theiet.org**

www.**theiet**.org/information-communications