Vol 1 | ISSN 2041-5923
2009

IET Journals

The Best of IET and IBC

INSIDE Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters.*

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

Contents

VOLUME 1 2009

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Introduction

Welcome to *The Best of IET and IBC*, a co-publication between the International Broadcasting Convention and the Institution of Engineering and Technology.

The IET is one of the members of the IBC partnership board and the two organisations maintain a close working relationship. This year we have come together to deliver something unique to the IBC conference, a new high-quality publication that will appeal to all those with a technical interest in the field of broadcast media. This publication contains, at a glance, the very best technical content from IBC 2009: a selection of papers from across the sessions, including the overall best paper, 'Compatibility Challenges for Broadcast Networks and White Space Devices', by Mark Waddell, and an interview with the best young professional poster author, Thomas Jaeger. This is complemented by specially selected technical journal content from *Electronics Letters*, the IET's flagship peer-reviewed journal.

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 spotlight this year is on young engineers and new enterprises – the creativity that we see being brought back into the marketplace and how innovation and enterprise are changing the landscape in the communications and media industries. The IBC 2009 conference committee has crafted an engaging programme of papers, panel discussions and masterclasses in response to a strong message from the industry that this is an exciting time for revolutionary technologies and evolving business models.

Collective inspiration

The Institution of Engineering and Technology, or IET for short, is one of the world's leading professional societies for the engineering and technology community. Not only does it have more than 150,000 members in 127 countries and offices in Europe, North America and Asia-Pacific, it is also the publisher of, among other things, a suite of 22 internationally renowned peer-reviewed publications that cover the entire spectrum of engineering and technology. Many of the innovative products on show at IBC will have been born from research published in IET titles such as *IET Image Processing and IET Computer Vision*. The papers you will read in this publication, however, are from the IET's flagship letters journal, *Electronics Letters*. Published in print every fortnight, it publishes cutting-edge technology papers faster

than any other printed journal, with many authors choosing to publish their preliminary results in *Electronics Letters* before presenting their results at conferences, such is the journal's reputation for quality and speed.

Working closely with the IET journals team are the IET Technical Professional Networks. These TPNs aspire to act as a natural home for people who share a common interest in a topic area (regardless of geography), foster a feeling of belonging to a community, and support two-way communication between network registrants, the IET and each other. Each network is assisted by an executive team, made up of willing volunteers who bring together unique experience and expertise for the benefit of network registrants. Members of the Multimedia Communications Network Executive Team have been instrumental in the creation of this publication, from reviewing articles to suggesting content. They have brought their industry know-how to the fore and helped show the key part volunteers have to play in widening the reach and influence of the IET.

On this note it only remains for me to extend my thanks to everyone involved in creating this special *Best of IET and IBC* publication. I am sure you will agree it reflects the engaging programme of papers, panel discussions and masterclasses taking place at this year's conference and the high quality of peer-reviewed research that you would expect from the IET. I hope that you enjoy it, and wish all of you attending this year an enjoyable IBC 2009.

Professor David Crawford
Chairman of the IBC Conference

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Editorial

IBC's New Technology Campus

Now in its 15th year, the New Technology Campus is a hub for international innovation where research organisations demonstrate their latest work to the IBC Conference and Exhibition delegates. The campus also allows the IBC conference authors to have an extended forum where they can demonstrate the technology on which they are speaking, and provides a platform to show the latest developments within the industry that are novel, relevant and of high interest. The over-riding requirement for would-be campus booth holders is that they should stage live, working demonstrations of their technology in an informal but informative way.

This year, once again we have an international gathering of researchers, representing **Japan, Canada and Europe**. In the main they represent the best and most interesting internationally collaborative projects.

The projects address a wide and diverse set of problems facing the media industry, and more importantly, offer an insight into potential solutions. Delegates who visit the campus booths will undoubtedly leave with new ideas that can really help their businesses, as well as an insight into future technology trends.

From **Europe** we already have the following eclectic mix of collaborative projects to display:

VITAL MIND is a project dedicated to supporting the elderly population and therefore includes the research and development of highly cognitive brain fitness content. Its objectives are the design and development of cognitive activities, the development of new methods for user control, the production of Authoring and Production Tools for iTV based cognitive training applications, and the promotion of the use of the USB Flash Device (UFD) as an addition to the broadcast delivery system.

The **3D4YOU** project is developing the key elements of a practical 3D television system, particularly the definition of a 3D delivery format and guidelines for a 3D content creation process. The project will show 3D capture techniques, conversion of captured content for broadcasting and the development of 3D coding for delivery via broadcast, i.e. suitable to transmit and make public. 3D broadcasting is potentially the next major step in home entertainment.

Continuing on the theme of 3DTV, **2020 3D Media** are researching, developing and demonstrating novel forms of compelling entertainment experiences based on new technologies for the capture, production, networked distribution and display of three-dimensional sound and images.

B21C (Broadcast for the 21th Century) is dealing with the specification, the verification or the validation of new standards for digital terrestrial TV and mobile TV. The project has the clear objectives of sustaining the evolutions of digital video broadcasting demanded by emerging new applications and usage of broadcast content.

The **ANSWER** project is showing techniques for Automated 3D Pre-Visualisation for Modern Production. Theirs is a new approach to the creative process of film and game production. Musicians and choreographers have long been able to express their intentions using logical symbolic structures (music notation and dance notation). ANSWER will produce a notation system for describing the creation of multimedia content, thus offering a bridge between digital media production and animation for game design.

SALERO (Semantic AudiovisuaL Entertainment Reusable Objects) aims at making cross media-production for games, movies and broadcast faster, better and cheaper by combining computer graphics, language technology and semantic web technologies as well as content based search and retrieval.

iNEM4U is short for interactive Networked Experiences in Multimedia for You. The aim of iNEM4U is a networked system that facilitates enhanced multimedia experiences for individuals and communities. Their objective is to research and develop a networked system that facilitates rich multimedia experiences across technology domains. These

experiences will be far better than today's experiences and will ultimately result in shared cross-domain experiences.

From **Canada**, the Communications Research Centre (**CRC**) will be presenting their Openmokast project, which is a framework for open source broadcast handhelds. CRC noticed that there is a strong enthusiasm for physical layer mobile broadcasting technologies such as DAB/DMB, DVB-H and MediaFLO in both the broadcast and telecom industries today. Lessons learned from the Internet ecosystem show that user innovation can be key in the creation of new and disruptive applications. The Openmokast project was launched by the CRC to catalyse application innovation in mobile digital broadcasting through the development of an open software stack and open digital broadcasting enabled handsets.

Finally from **Japan** we have contributions from NICT and NHK.

The National Institute of Information and Communications Technology (**NICT**), which is the government-funded telecommunications research organisation in Japan, are presenting a demonstration of their research on the human interface to communications media. Their Multisensory Interaction System is one where users can feel the presence of a virtual object, enhanced by a combination of visual, haptic and auditory clues.

In contrast, the Japanese public broadcaster **NHK** is showing their Advanced Interactive Broadcasting Service Platform using Java data broadcasting. Since digital broadcasting launched in Japan in 2000, data broadcasting has been serving a wide variety of information, such as news, weather, traffic information as well as program related data broadcasting. NHK are therefore developing a new type of data broadcasting as an advanced service platform that is capable of providing services to satisfy the recently diversified needs of viewers.

Whether your interest is in new consumer interface mechanisms, mobile service enhancements, 3D television or awareness of the needs of an aging population, you will find the answers and knowledgeable research teams on the New Technology Campus at IBC 2009.

David Meares
Organiser of the New Technology Campus

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Compatibility challenges for broadcast networks and white space devices

M.B. Waddell

BBC, Surrey, KT20 6NP, UK
E-mail: mark.waddell@bbc.co.uk

Abstract: Digital switch-over plans have driven a thorough review of ultra high-frequency spectrums and how they might be used. The deployment of low-power white space devices (WSDs) has the potential to deliver improved WiFi systems for mobile broadband and a new platform for multimedia streaming in the home. The FCC has recently approved plans for new fixed and mobile devices, based on a combination of spectrum sensing and geolocation. Devices are expected to appear in the market in the very near future, but will require significant modification to cope with the denser, higher-value network of transmitters in Europe. The spectrum sensing approach is at an early stage of development and in isolation is unlikely to provide the level of protection required to prevent interference to broadcast services and radio microphones. However, when this technique is combined with Geolocation techniques with EIRP control, the technology rapidly becomes viable. The potential opportunity of a new harmonised, licence-exempt band to support on-demand multi-media streaming in the home is an irresistible target and the technology is creating considerable interest.

1 Introduction

Ultra high-frequency (UHF) terrestrial TV networks have historically been planned as multi-frequency networks to support regional TV programming and to simplify international frequency co-ordination. This can be seen as a relatively inefficient use of the spectrum as a particular UHF channel carrying a TV multiplex for one region cannot be re-used until the signal strength has fallen to a level approaching the thermal noise floor. In the UK, 256 MHz of spectrum is used to support six DTT multiplexes, each 8 MHz wide. At any particular location, there will be a significant number of 'empty' channels that cannot be used for additional high-power TV services without causing interference to services in adjacent regions. Traditionally, these channels, known as UHF white space, have been used for low-power applications in programme making and special events (PMSE), typically radio microphones and wireless in-ear monitors (IEMs). This usage is fairly sparse, however, and the possibility of using the white space for new, low power, licence-exempt devices would provide an additional, much-needed band to supplement the popular but crowded 2.4 GHz ISM band. This could potentially support high bandwidth wireless applications like multimedia streaming, video on demand and TV catch-up services, which would be of particular interest to broadcasters.

2 White space access

Accessing the UHF white space for unlicensed applications has proved quite controversial, with existing licensees understandably nervous about the risk of interference to their services. Since the channel availability varies across the country, assigning white space allocations and access is not straightforward. To address interference concerns, two techniques are emerging for UHF white space access: spectrum sensing and geolocation.

2.1 Spectrum sensing – 'cognitive access'

The simplest access approach is to scan the TV spectrum for an unused channel and use this for the white space application on a 'listen and broadcast when clear' basis. This is superficially very attractive: it requires no additional hardware or infrastructure as the white space device (WSD) can make use of the tuner and antenna needed for its own applications to carry out the initial spectrum scan.

Unfortunately, the process is difficult and requires high-performance radio frequency (RF) circuitry and potentially complex signal processing as the licensed DTT signal to be detected will be received at a very low level. The WSD will be lower in height than a normal DTT antenna, it will have a lower antenna gain and will have an obstructed view of the transmitter. These effects combine to give a hidden node margin; this hidden node margin relates the rooftop antenna signal level for DTT reception to the WSD signal level available for detection. The components of the hidden node margin are shown in Fig. 1.

Research by Randhawa *et al.* [1] suggests that for outdoor suburban deployments of WSDs in the UK, the hidden node margin will be as high as 40 dB. This figure is based on outdoor sensing at 1.5 m with a 0 dBi antenna. Assuming a planned field-strength of 50 dB μV/m at 10 m, which would typically deliver −72 dBm to a DTT set top box, the required detection sensitivity for a WSD would be −112 dBm. For indoor deployment of WSDs, where building penetration and reflections further reduce signal level, the available signal strengths will typically be 20 dB lower, suggesting an indoor hidden node margin of 60 dB.

The difficulty of detecting such small signals can be appreciated by considering the signal-to-noise ratio available for sensing by the WSD. Using typical TV planning parameters taken from the Chester 97 DTT planning agreement [2], Table 1 shows how the available signal-to-noise ratio is degraded from the TV planning value by location, antenna gain and EMI effects. Detection of DTT signals buried in noise will be exceedingly difficult.

A number of prototype devices were assessed by the FCC in 2008 and the results of these tests by Jones *et al.* [3] demonstrated the detection of clean ATSC DTV signals at −116 dBm in a 6 MHz channel. This sensitivity is impressive, but unfortunately detection performance degraded significantly in the presence of high (−28 dBm) and moderate (−53 dBm) level signals in the adjacent and alternate channels. Some devices malfunctioned completely,

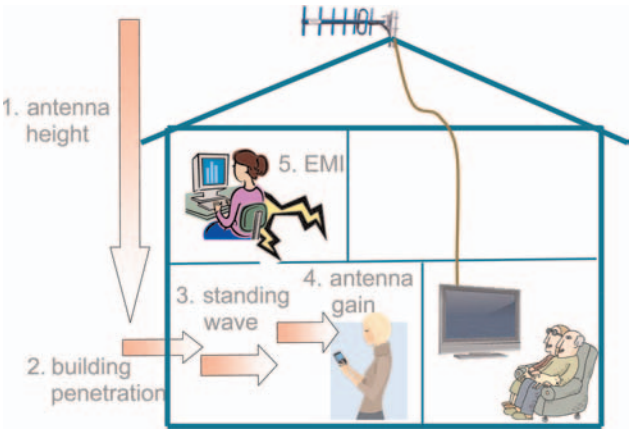

Figure 1 *Hidden node margin loss components*

Table 1 Carrier to noise ratio for DTT detection indoors at 1.5 m

Required DTT CNR for QEF (64-QAM rate 2/3), dB		19
planning margin, dB		+8
DTT antenna gain, dBi	12	
WSD antenna gain, dBi	−10	
C/N loss at WSD antenna, dB		−22
height loss, dB	12	
building penetration loss, dB	7	
indoor location variation (95%), dB	14	
C/N losses due to location, dB		−33
degradation due to EMI, dB		−8
S/N at WSD, dB		−36

while others were desensitised by up to 60 dB, indicating device dynamic range limitations.

To achieve their sensitivity, the US prototype devices have exploited the pilot tone in the ATSC DTV signal, which can be clearly seen on a spectrum analyser plot (Fig. 2). For detection of DTV at −116 dBm an overall signal-to-noise ratio of 13 dB is available assuming a 3 dB noise figure. However, by using a simple bandpass filter centred on the pilot-tone, the signal-to-noise ratio improves significantly; for a 1 kHz bandwidth filter, a signal-to-noise ratio of 13 dB becomes available, which is more than sufficient to rapidly detect the pilot.

Detection of COFDM systems like DVB-T and DVB-T2 will require far more sophisticated signal processing using correlation of the guard interval or detection of the OFDM pilot structure. This process is further complicated by the multiplicity of modes and has not yet been demonstrated on practical devices.

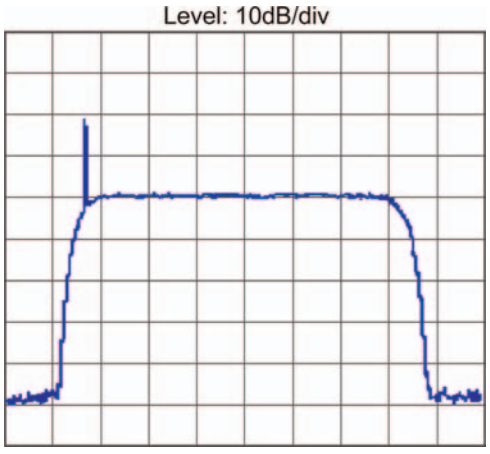

Figure 2 *ATSC DTV spectrum showing Pilot tone at −11.3 dB*

2.2 Geolocation

An alternative to sensing is to control access using an Internet-hosted, location-dependent database of available white space channels. A device would typically use GPS to locate itself and then request a table of available channels from a server. This avoids the difficulties associated with detection but clearly requires some additional hardware and infrastructure.

This technique is particularly appropriate for DTT protection, where channel assignments are essentially static but can readily be extended to protect PMSE use where access is licensed and logged by a band manager. This is particularly attractive in the UK where the existing PMSE band manager already makes extensive use of computer databases to licence radio microphone users. In some countries PMSE is less well controlled and sensing techniques may still be necessary. PMSE sensing performance issues remain a concern and the FCC have chosen to adopt a 'safe haven' approach reserving two location-dependent TV channels for exclusive PMSE use.

3 WSD EIRP limits

Access to the white space channels using geolocation techniques should prevent co-channel interference, but careful control of the EIRP will be needed to prevent adjacent and non-adjacent channel interference. Ideally, devices will make use of power control to minimise interference and maximise opportunities for spectrum re-use. However, sensible EIRP limits will be required to protect licensed incumbents and these must take account of typical antenna isolation values and the selectivity and overload characteristics of the existing receivers.

3.1 WSD to DTT receiver path loss

The path loss between the WSD and the DTT receiver is clearly a crucial factor. Initial analysis by Ofcom [4] considered a WSD outdoors at 1.5 m height, 45° off axis to the DTT antenna as shown in Fig. 3. The minimum distance between the WSD and DTT antenna would be 10 m, corresponding to a free space loss of 50 dB at 800 MHz. The WSD is off axis to the DTT antenna and was assumed to be 10 dB down in gain from boresight, that is 2 dBi, and the WSD antenna was assumed to be 2 dB down from its peak value, that is −2 dBi. For 800 MHz operation, a feeder loss of 5 dB was assumed, giving a total path loss of $50 + 5 + 2 - 2 = 55$ dB.

Ofcom's initial analysis may slightly overestimate the path loss for some deployment scenarios however, for example at the lower end of the UHF band or for indoor use. This is of potential concern as the resulting WSD EIRP proposal might still cause interference to some DTT receiver installations. For example, at 500 MHz, a 2 dB feeder loss

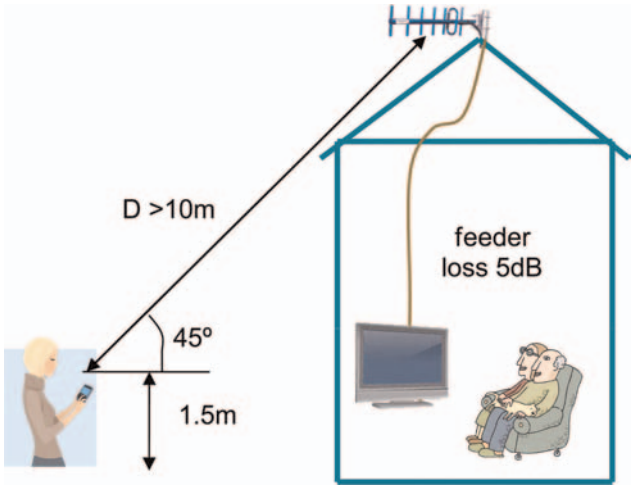

Figure 3 *Estimation of outdoor WSD to DTT receiver path loss*

would apply and the free space path loss would reduce to 46 dB giving a total path loss and EIRP recommendation 7 dB lower than the Ofcom figure.

Indoor deployments are the most challenging and protecting portable receivers or DTT loft antenna installations may prove very difficult. Fig. 4 shows a WSD access point in the loft space of a semidetached property. This could be less than 5 m from a loft mounted DTT antenna in the adjacent property. Assuming a 7 dB building penetration loss between buildings and on-axis coupling between the WSD and DTT antennas, the path loss at 500 MHz would be 18 dB lower than the Ofcom figure. This analysis is summarised in Table 2.

3.2 DTT receiver selectivity – C/I performance

Receiver C/I performance is a measure of selectivity defining the permitted level of interference for a given signal level and frequency offset. Performance depends on the DTT mode

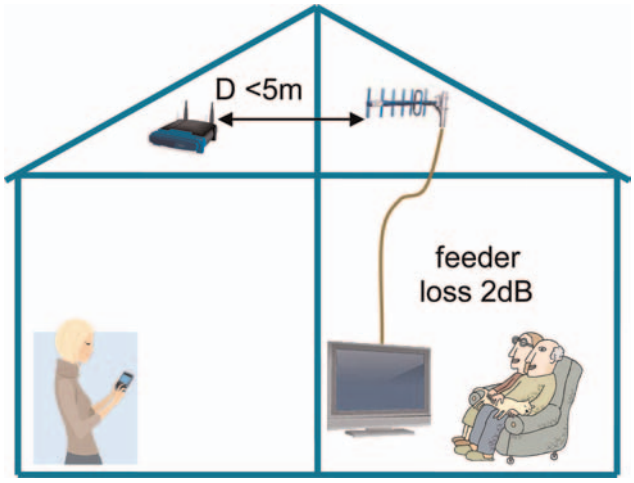

Figure 4 *WSD to DTT receiver path loss for loft installations*

Table 2 WSD to DTT receiver path loss scenarios

Scenario	1. Outdoor band V	2. Outdoor band IV	3. Adjacent lofts
geometry	Fig. 3	Fig. 3	Fig. 4
frequency, MHz	800	500	500
distance (D), m	10	10	5
free space loss (l), dB	50	46	40
DTT antenna gain (r), dBi	2	2	12
WSD antenna gain (w), dBi	−2	−2	0
feeder loss (f), dB	5	2	2
building penetration loss (b), dB	0	0	7
path loss ($l - r - w + f + b$), dB	55	48	37

and signal level and the interferer frequency offset. At low DTT signal levels the permitted interference level will be noise limited while at higher levels, non-linear effects become apparent and performance degrades again. For a typical receiver there is a region where the permitted C/I level is constant and this can be used for the EIRP limit calculations.

It is too early to characterise the actual interference rejection performance of DTT (or PMSE) receivers to WSD interference, as the characteristics of the WSD signal are yet to be defined. An estimate of the likely performance can be made by assuming the WSD signal will be noise-like and similar to a DTT signal in its spectrum. The performance of real DTT receivers has been characterised in great detail and target specifications are published in the DTG-D book [5]. These performance targets are summarised in Table 3.

3.3 Estimate of WSD EIRP limit

By considering the minimum path loss between the WSD and the victim DTT receiver, the C/I performance of the receiver and the planned field strength, the maximum permitted EIRP for the WSD can be estimated for different scenarios. In the UK, the country is split into 100 m by 100 m squares, or pixels, and the DTT field strength is planned to exceed a mean value of 50 dB μV/m, for 99.9% of these pixels. Assuming a DTT antenna gain of 12 dBi, the received power can be calculated and is shown in Table 4.

Using the received power values from Table 4 and the DTG C/I targets in Table 3, the permitted WSD EIRP for each reception scenario can be predicted and is shown in Table 5.

Note that the EIRP limits are somewhat smaller than the values initially suggested by Ofcom (+20 dBm non-adjacent, 13 dBm adjacent) for a number of reasons. Ofcom have assumed receivers will outperform the DTG C/I performance targets by 7 dB and have allowed an additional 3 dB feeder loss in their Band IV analysis than that usually used for TV planning. Loft installations were not considered and may be very difficult to protect.

Table 3 DTT receiver C/I targets

Test condition (offset)	Interference level, dBm	C/I target (QEF), dB
ACI ($N \pm 1$)	−25	−27
non-ACI ($N \pm 2$)	−25	−38
non-ACI ($N \pm 3$)	−25	−43
non-ACI ($N \pm$ M, $M > 4$, $M \neq 9$)	−25	−47
non-ACI ($N + 9$)	−25	−31
linearity test (two interferers at $N + 2$, $N + 4$)	−25	−28

Table 4 DTT signal levels for outdoor and loft reception

Scenario	1. Outdoor band V	2. Outdoor band IV	3. Adjacent lofts band IV
field-strength at 10 m, dB μV/m	50	50	50
antenna gain, dBi	12	12	12
antenna shielding, dB	0	0	7
feeder loss, dB	5	2	2
received power, dBm	−78	−72	−79

Table 5 WSD EIRP limits

Deployment scenario	Test condition (frequency offset and receiver C/I)				
	Adjacent ($N \pm 1$)	Alternate ($N \pm 2$)	Non-adjacent ($N \pm M$, $M > 4$, $M \neq 9$)	Non-adjacent ($N = 9$)	Linearity limited ($N + 2$, $N + 4$)
	C/I = −27 dB	C/I = −38 dB	C/I = −43 dB	C/I = −31 dB	C/I = −28 dB
1. outdoor band V	4 dBm	15 dBm	20 dBm	8 dBm	5 dBm
2. outdoor band IV	3 dBm	14 dBm	19 dBm	7 dBm	4 dBm
3. adjacent lofts band IV	−15 dBm	−4 dBm	1 dBm	−11 dBm	−14 dBm

This analysis does not take account of location variations associated with the log-normal variation of field-strength within a planning pixel and this could reduce received power and EIRP still further. Furthermore, the use of domestic low noise amplifiers for signal distribution has not been considered and this can result in premature overload and degraded C/I performance.

4 Conclusions

Licence-exempt use of the UHF white space will become increasingly important as an alternative to the congested 2.4 GHz ISM band for low-power, broadband and multimedia applications. The opportunity of an internationally harmonised, licence-exempt spectrum band is so attractive that the development of devices seems inevitable.

Cognitive spectrum sensing is attractive in principle, but the sensitivity, RF dynamic range and signal processing requirements are beyond that which can be reliably achieved with current technology. Requirements for outdoor sensing are difficult enough and indoor sensing looks virtually impossible. OFDM signals are far more difficult to detect than the ATSC signals so sensing may prove particularly impractical in countries using DVB-T and DVB-T2.

Geolocation is emerging as the preferred technique and WSDs will require GPS or similar location capability and Internet access to access the channel tables. This will prevent co-channel interference to incumbent PMSE and DTT, but adjacent channel interference remains a concern. Worst-case adjacent-channel inference analysis suggests that indoor DTT installations, including portable and loft mounted antennas, may be particularly vulnerable to interference.

To control adjacent channel interference it is desirable to extend the database to include EIRP values for each of the available white space channels. The EIRP will be a function of the level of the neighbouring licensed services and the performance of the receiver. Locations at the edge of TV coverage will require lower EIRP limits than those enjoying increased coverage margins. Improved receiver performance may allow increased EIRP in the future. By including EIRP in the geolocation databases, device limits can evolve with time as understanding of the interference problems improves.

5 Acknowledgments

The author thanks his colleagues at the BBC for their contributions to this work. He also thanks the BBC for permission to publish this paper.

6 References

[1] RANDHAWA B.S., WANG Z., PARKER I.: 'Analysis of hidden node margins for cognitive radio devices potentially using DTT and PMSE spectrum', http://www.ofcom.org.uk/radiocomms/ddr/documents/eracog.pdf, 2009

[2] European Conference of Postal and Telecommunications Administrations, Chester, 25 July 1997: 'The Chester 1997 Multilateral Coordination Agreement relating to Technical Criteria, Coordination Principles and Procedures for the introduction of Terrestrial Digital Video Broadcasting (DVB-T)', http://www.ero.dk/132D67A4-8815-48CB-B482-903844887DE3

[3] JONES S.K., PHILLIPS T.W., VAN TUYL H.L., WELLER R.D.: 'Evaluation of the performance of prototype TV-band white space devices phase II'. OET Report FCC/OET 08-TR-1005, http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-08-2243A3.pdf, 2008

[4] Ofcom: 'Digital dividend: cognitive access-consultation on licence-exempting cognitive devices using interleaved spectrum', http://www.ofcom.org.uk/consult/condocs/cognitive/cognitive.pdf, 2009

[5] DTG: 'D book requirements for interoperability V5.02', 2008

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Stereoscopic three-dimensional sports content without stereo rigs

O. Grau V. Vinayagamoorthy

BBC R&D, UK
E-mail: oliver.grau@bbc.co.uk

Abstract: An alternative approach to generate stereoscopic content of sports scenes from regular broadcast cameras without the need for special stereo rigs is contributed here. This is achieved by using three-dimensional (3D) reconstruction previously developed for applications in post-match analysis. The reconstruction method requires at least four to five cameras and computes the 3D information automatically. Two different target formats for the delivery of stereoscopic 3DTV are discussed: the display-independent layered depth video (LDV) format and conventional binocular stereo. The results viewed on two different displays demonstrate the potential of the method as an alternative production method for stereoscopic 3D content.

1 Introduction

Owing to the successful re-introduction of stereoscopic 3D (S3D) in the cinemas, there is recently also a growing interest in the broadcast industry in S3D content. In particular, the coverage of live sports events like football or rugby is a focus of interest. Unfortunately, the production of stereo content increases costs significantly. To add S3D to the regular broadcast each camera position would need to be equipped with a pair of cameras in a special stereo rig. That would increase the costs beyond simply doubling the camera budget, since it requires special skills to set up, to operate and additional broadcast infrastructure, for example, in outside broadcast (OB) vans.

At the current state, S3D content is captured with special stereo rigs, which are either built up using miniature cameras or through the use of a special mirror rig, so that the content can be captured with an inter-ocular distance of approximately 6.5 cm, which represents the average eye distance of the population. Special camera systems, like boxed super-zoom-lenses or high-speed cameras that are commonly used for the coverage of sports content cannot be easily used in a stereo rig. For this and other reasons, it is very likely that S3D productions will only be able to share some of the conventional broadcast equipment and will be operated mostly alongside the conventional broadcast coverage.

In this contribution, we present an alternative production method, which derives 3D stereoscopic content from the normal broadcast coverage cameras, plus optional locked-off cameras. The approach makes use of multi-camera 3D reconstruction techniques that have been previously developed for post-match analysis of sports scenes [1, 2]. These techniques perform an automatic camera calibration and segmentation of the foreground action. From this information, a 3D model of the foreground action is computed. These data are then converted into a 3DTV format. The '3D4You' project [3] investigates formats that are independent from the 3D display as it stores depth information alongside the image, allowing the generation of either stereoscopic or multi-view for different kinds of 3D displays. Alternatively, the data can be stored as a conventional S3D image pair for the left and right eye with a fixed inter-ocular distance.

The rest of this paper is structured as follows. The next section gives a very brief overview of our approach, followed by a more detailed description of the processing modules. The section following that discusses 3DTV delivery formats and aspects of the method that has been developed to convert the 3D action into these formats. The paper finishes with a description of the first results and some conclusions.

2 Overview of the system

As depicted in Fig. 1 the system has three component blocks:

1. The capture system is fully integrated into the OB infrastructure and performs a live ingest of the multi-camera

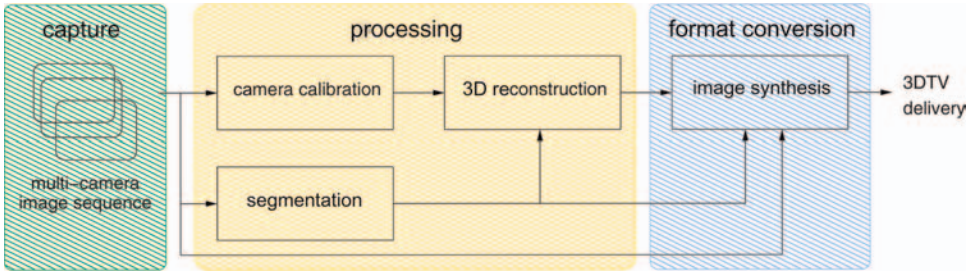

Figure 1 *System overview*

video streams. For our tests, we implemented the capture system with standard IT server hardware equipped with HD-SDI frame grabber cards.

2. The processing block automatically computes 3D information from the multi-camera streams.

3. The format conversion converts the image- and 3D-data into a 3DTV delivery format.

In the current experimental implementation, the video streams are stored to disk whereas the processing and format conversion are run offline to produce the stereoscopic 3DTV content.

3 Processing modules

3.1 Camera calibration

The camera calibration determines the parameters of the broadcast cameras. Our approach does this automatically from the images without any additional special hardware (like pan-tilt heads). The method detects the pitch lines of a sports ground, the dimensions of which need to be known. From this information, the camera position, orientation and internal parameters (focal length and lens distortions) are determined. A detailed description of the calibration method can be found in [4].

3.2 Foreground segmentation

The foreground segmentation separates the foreground action from the background, that is, the pitch and other background objects like stands.

For the segmentation of players, colour-based methods such as chroma-keying against the green of football and rugby pitches have been considered. However, the colour of grass varies significantly on pitches. This is owing to uneven illumination and anisotropic effects in the grass caused by the process of lawn-mowing in alternating directions. Under these conditions, chroma-key gives a segmentation that is too noisy to achieve a high-quality visual scene reconstruction. Therefore two improved methods have been implemented and tested: a global colour-based '*k*-nearest-neighbour classifier' classifier (KNN) and a motion compensated difference-keyer. Details of the KNN classifier can be found in [2].

The difference-keyer derives a key based on the difference in colour of a pixel in a camera image to a model of the background that is stored as a background plate. This method is usually applied to static cameras. However, under known nodal movement of the camera, a background plate can be constructed by piecewise projection of the camera images into a spherical map. This transformation is derived from the camera parameters, as computed in the camera calibration. A plate clear of foreground objects is created by applying a temporal median filter to the contributing patches. An example background plate is shown in Fig. 2. The actual difference key is then derived from the difference of a pixel in the current camera image to its reference in the spherical map, taking into account the camera movement and zoom. The latter parameters are known from the camera calibration.

The difference-keyer method is superior in most cases to a global colour-based approach. It is even able to separate between foreground and mostly stationary but cluttered backgrounds and is used for the results presented in this

Figure 2 *Spherical background plate of a rugby pitch*

paper. The colour-based approach on the other hand is relatively simple to implement, but limited mainly to the segmentation of players against the pitch. Fig. 3 shows an example difference key for a moving camera.

3.3 3D reconstruction of foreground action

The 3D reconstruction uses a visual hull or shape-from-silhouette approach to compute 3D models of the scene action (for more details, see [2]). For the rugby scene shown in Fig. 3, this is done for an area of $50 \times 50 \times 3$ m (width $\times$ depth $\times$ height) with a volumetric resolution of $512 \times 512 \times 16$. This only gives a relatively coarse 3D surface model, but the image rendering makes use of the keys generated by the segmentation and uses them as an alpha channel to obtain a better shape of the boundaries. Fig. 4 shows a 3D model overlaid onto the original camera image.

A limitation of using only 3 m height is that the ball will not be represented when it is kicked higher than this. Using a higher volumetric area is not without problems. One problem is the increase in the computational effort required. Another problem is that the only areas of the active volume that can be reconstructed are those seen by at least one camera. Since the 'higher' areas are only marginally of interest during most of a game these areas would not have been covered by many cameras. The approach taken to tackle this in the 'iview' project was to model the ball in a separate pass: an operator sets the active area manually to roughly where the ball is. Although this is acceptable for post-match analysis, it would be desirable to automate this approach for S3D coverage.

3.4 3D reconstruction of background objects

The generation of stereoscopic images also requires the 3D geometry of background objects. The pitch can be approximated by a planar polygon, as its dimensions are known, as a pre-requisite for the camera calibration step. In comparison to the foreground action, the stadium is further away and for our initial tests we approximated it with very

Figure 4 *Wire-frame overlaid 3D model*

few polygons. This can be done with a CAD or post-production modelling tool. If more detailed models are needed then these have to be aligned to the images. This can be achieved with image-based modelling tools, which is a subject of our current investigations.

The modelling of the background objects can be done offline in preparation of a game and needs to be done only once per location.

4 Conversion to stereoscopic format

After the reconstruction of foreground action and background, these objects are made available as 3D descriptions in the form of a polygonal surface mesh. This scene description is then the basis for the conversions into a 3DTV format.

There is currently no regular 3DTV service available and moreover there is, as of yet, no standard format. A likely candidate is a side-by-side aligned stereo image pair (binocular stereo). One implementation could be to scale the left and right image 50% horizontally and merge them into a new image, which is of the same size as the original ones. This will reduce the horizontal resolution of the images, but demands no further changes in the rest of the transmission chain other than a display capable of handling this format on the viewer side. The viewer cannot change

Figure 3 *Camera image (left) and difference key (right)*

the depth scaling at his end as the inter-ocular distance is fixed with this format.

Alternative approaches for 3DTV delivery are investigated in the 3D4You project [3], with the goal to define a display-independent delivery format. One option is to use one video stream plus depth information (stored like an alpha-channel). An extension of this 'video-plus-depth' is a 3D-layered depth video (LDV), which is constructed by adding additional occlusion layers. The end product is one central camera view with all video and depth data from multi-view capturing mapped during post-production. This format is very compact and efficient in terms of data compression and low-complexity rendering on a 3D display.

The conversion of the 3D data acquired by our approach into the LDV or conventional binocular stereo formats is achieved by synthesising the required information from the reconstructed 3D scene description.

4.1 Conversion to LDV

To generate the LDV format, the original camera images are augmented by a depth channel and by a background layer. The resulting data were viewed on a Philips WOWvx display [5]. The depth information was generated by rendering the 3D scene description with a scan-line renderer. Our implementation is based on OpenGL. The depth information is retrieved by reading out the z-buffer.

The depth values in LDV (and image-plus-depth) are mapped to byte-integers. The depth range is therefore limited to a minimum and maximum depth value in a particular scene. The depth values are converted to disparity values within a range of 0–255, where a value of 0 corresponds to objects located at the maximum disparity behind the screen while 255 corresponded to objects closest to the observer [5]. A disparity value of 128 corresponded to objects on the screen. The disparity values of objects such as the goal posts in Fig. 6 were constructed with disparity values between 128 and 255 in order to project them as foreground objects to the observer.

The background layer is an image from the same camera angle as the camera image, but without the foreground action. One option to generate this layer is to fill in the obscured background from information taken from other cameras. One problem found with this procedure is that the areas might look slightly different in colour because of different colour characteristics of the cameras (mismatched colour balance) or anisotropic effects – see remarks about segmentation above. Furthermore, the occluded area might not be visible in other cameras with the same degree of detail.

Another approach to fill un-revealed areas is to use background images generated from the spherical background plate as used in the difference keying technique described above. Although the generated images are slightly less detailed and miss shadows, the results of this approach look quite promising. Fig. 5 shows the generated background layer for the camera angle of the scene depicted in Fig. 3.

Figure 5 *Background layer*

The resulting 2D, depth and background layer sub-images are merged into one image in the required format to provide a 1920 × 1080 image for each frame to be played on a 42 inch Philips WOWvx display [5].

4.2 Conversion to binocular stereo

For binocular stereo the original, monoscopic camera view is used as the left view and the image for the right view is synthesised by rendering from a camera viewpoint that is laterally offset by the intra-ocular distance. An advantage of this approach is that the intra-ocular distance can be freely varied. This is of particular interest since the optimal distance depends on the screen size and viewing distance. Screens in the cinema usually require a different inter-ocular distance than for domestic TV viewing.

5 Results

Three short sequences, with a total length of about a minute, taken from a pre-recorded rugby game have been used to test the approach. A number of broadcast cameras from that game were captured and four camera positions were then used. A further six additional locked-off cameras were recorded to complement the broadcast cameras. From this set of 10 cameras, a 3D model of the action was computed.

The 3D data set was then converted into image-plus-depth, LDV and binocular stereo. Fig. 6 shows an example frame from that data set from different camera angles. The image-plus-depth and LDV were displayed on an auto-stereoscopic Philips WOWvx display. The depth augmentation works very well, resulting in a good visual quality. The lack of detail in the background objects (stadium) is not disturbing.

The binocular data were viewed on a 19 inch Trimon ZM-M190 display using polarisation glasses. Compared to the Philips display, the image resolution is higher and artefacts

Figure 6 *Images from a rugby game (left) and synthesised depth (right)*

are more clearly visible. In particular, it would be desirable to have finer detail in the background model as it can be seen that parts are too flat.

6 Conclusions

This paper describes an alternative approach to generate stereoscopic content of sports scenes from regular broadcast cameras. This is achieved by using 3D reconstruction previously developed for applications in post-match analysis and synthesis of the target 3DTV format. The reconstruction method needs at least four to five cameras to generate the 3D information automatically. This number of cameras is usually available in the coverage of high-end games. In order to improve the robustness and quality of the system, additional locked-off cameras can be used.

The availability of a 3D description of the scene has the major advantage that different target formats can be handled. That includes the display-independent delivery formats discussed, like LDV, or the ability to render binocular stereo sequences with different, fixed inter-ocular distances for different end-terminals.

Further research is currently being carried out to make the methods more robust so that the processing would run fully automated. Although the experimental implementation of the system currently runs offline, it is believed that the algorithms used could be optimised and run in real time.

7 Acknowledgments

This research was supported by the TSB project iview and FP7 3D4You.

8 References

[1] The iview project, http://www.bbc.co.uk/rd/projects/iview

[2] GRAU O., THOMAS G.A., HILTON A., KILNER J., STARCK J.: 'A robust free-viewpoint video system for sport scenes'. Proc. 3DTV Conf., Kos, 2007

[3] IST 3D4You project, http://www.3d4you.eu

[4] THOMAS G.A.: 'Real-time camera tracking using sports pitch markings', *J. Real Time Image Process.*, 2007, **2**, (2–3), pp. 117–132

[5] 3D interface specifications of the WOWvx display, White paper Philips 3D solutions, February 2008

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronic Letters*

ISSN 2041-5923

Keeping the HD quality high for events using standards converters and new HD contribution codecs

A. Kouadio

EBU Technical, 1218, Switzerland
E-mail: kouadio@ebu.ch

Abstract: It is no longer abnormal for large events such as the Olympic Games to be shot in HDTV. With the migration of many broadcasters towards HDTV, HDTV signals need to be transmitted around the world, and need to be standards' converted, while maintaining sufficient quality headroom for final broadcasting. Outlined is the results of comprehensive EBU tests on different HDTV standard converters in conjunction with contribution codecs using state-of-the art MPEG-2 compression and newly introduced H.264/AVC and JPEG2000 compression codecs. In addition, the study will give preliminary guidance on the impact of the position of the standards converter in the signal contribution chain. The study of the new contribution codecs examines their coding gain with respect to legacy MPEG-2 systems. The result of these ongoing investigations will be presented during the session at IBC 2009.

1 Introduction

HDTV broadcast quality depends on several factors throughout the supply chain (from the cameras to the end user display). Several studies have been made investigating production codec issues in order to provide EBU members with neutral and objective guidance for optimal technology choice. A substantial amount of today's HDTV content is programmes exchanged among broadcasters including from the US. For Europe, there is particular interest in international events such as the Olympic Games (i.e. winter games in Vancouver). Thus, the HDTV content generated at the event venue need to be 'contributed' to the broadcaster's location. In the case of an event outside Europe and produced in a different HDTV standard than that used by the broadcasting station (e.g. produced in 720p/60 or 1080i/30 and broadcast by a European broadcasters, and thus in 50 Hz standards) HDTV standards conversion is required too.

Contribution can be defined as any application where the signal is not delivered directly to the end user and where some post-processing may occur before the signal reaches the final viewer. Different networks can be used for the content delivery (satellite, fibre, radio links, SDI etc.) each having its own advantages and constraints. Since the introduction of digital SDTV and HDTV in Europe, the MPEG-2 4:2:2 compression system has been the de-facto standard used for contribution applications. However, recently other compression systems such as H.264/AVC, JPEG2000 or DIRAC have been proposed as serious candidates to replace MPEG-2-based contribution codecs. The main motivation is to reduce bit-rate requirements on the one hand, and to improve the image quality and delay after transmission on the other hand.

In the following, the relevant new compression systems are briefly outlined.

- *MPEG-4 AVC/H.264*

The successor to MPEG-2 in digital TV distribution standards was MPEG-4 Part 10 also known as H.264/MPEG-4 AVC [1]. It is the result of a joint standardisation effort (joint video team) between the MPEG and the ITU-T. The compression algorithm is widely used in multimedia applications and well accepted as distribution format for the consumer and in the professional industry.

H.264/AVC is supported by a comprehensive toolbox providing additional functionalities such as CABAC statistical coding, smaller macro block sizes, de-noising filters and others. H.264/AVC provides a significant coding gain of about 50% for distribution applications compared to former MPEG-2 systems (EBU BPN 076-081). However, the gain in bit rate for professional applications such as contribution with the relevant coding algorithms (e.g. 4 : 2:2 sampling) was thought to be more modest, but has never been assessed in detail. The following analysis will try to clarify this point. A good reference on H.264/AVC can be found in [2, 3].

• *JPEG2000*

This coding technology from the JPEG committee was intended to replace the famous and well-known still image standard JPEG. Owing to its intrinsic computational complexity, the JPEG2000 standard [4] was not as embraced by the still image camera industry as its predecessor JPEG had been. However, with the constant improvement in computing technology, the introduction of HDTV and digital cinema, it is progressively finding a new home in the video domain.

There are plenty of reasons for its adoption by the professional video industry. JPEG2000 is based on a different transform than used in the MPEG family (wavelet based instead of DCT) preventing blocky artefacts which are commonly found in DCT-based MPEG systems. The transform is applied as a succession of low-pass and high-pass filtering process. When they occur, JPEG2000 artefacts are rather blurry type, which is less annoying to the human visual system.

As in a still image codec, it is used as an intra-frame coding system, which increases its robustness to cascading. The JPEG2000 bit streams are highly scalable (spatial, temporal and quality scalability) providing it with robustness to transmission error as well as graceful degradation features. Being an intra-frame codec it requires very high bit rates compared to group-of-picture-based compressions systems such as H.264/AVC. It is thus mainly applicable to high bandwidth contribution links (fibre, radio links). More information on JPEG2000 can be found in [5, 6].

• *Dirac*

Dirac is an open source compression system developed by the BBC. It is based on wavelet technology as JPEG2000 and uses wavelet filters of the same family (Daubechies and LeGall). However, it also associates motion compensation to the wavelet transform to improve the compression capabilities and sustain high image quality at a lower bit rate. Dirac exists for professional applications (Dirac Pro 1.5 – HD-SDI compatible and Dirac pro 270 – SDI compatible) such as for contribution, where it is intra-coded only like JPEG2000 to sustain very low coding delays. It is also standardised by SMPTE as VC2. Since its hardware version has a fixed output bit rate, Dirac was not included in the test described by this paper.

During the contribution of international events between broadcaster's facilities, it is often required to perform standards conversion. After the impact of the contribution compression system, standards conversion is the second main quality factor in a contribution environment. Standard conversion is the process of adapting one TV standard to another. This process may require spatial and/or temporal conversions.

Standard conversion has become more complex with the introduction of HDTV, because of additional HDTV image formats and frame rates as compared to the existing SDTV formats. Spatial conversion and more specifically the interlace-to-progressive scan conversion process might require efficient de-interlacing algorithms (motion compensated) to maintain a satisfactory level of vertical resolution. The most complex operation still remains frame rate conversion. Temporal conversion can be done according to the following practices:

1. Simple linear inter-field interpolation.

2. *The motion adaptive multi-field interpolation:* which adapts its conversion strategy according to the amount of motion detected in the scene.

3. *Motion compensation:* which aims at removing the possible temporal conversion judder. Efficient motion compensation relies on efficient motion estimation. Several techniques are described in the following section:

• *Motion estimation:* Motion estimation is the process of determining positions with certain accuracy from one frame/field to another.

• *Block matching:* A specific block of pixel in an image is compared to a block of pixels at the same spatial location in the next image. Depending on the level of correlation between the block pixels, motion can be detected. In the case of a low correlation value, the block will be correlated to other blocks of the same size in a predefined search area. The block with the highest correlation is assumed to be the final destination.

• *Gradient method:* The gradient method tries to estimate motion by tracking a particular variation (gradient of luminance versus position) of luminance across an image coupled with the temporal variation of luminance (gradient of luminance against time). This technique can lead to false detection, for example when an object luminance changes while in motion (shading etc.) or if a similar luminance gradient happens to be in the same image at a different position.

• *Phase correlation:* The phase correlation is an efficient motion estimation process. It works by a spectral analysis of the search area in the two successive fields. The correlation provides peaks, the highest representing the object displacement. Phase correlation is a very accurate and robust (to noise, luminance variation etc.) motion estimation technique since all processing is done in the frequency domain.

2 Experiments

2.1 Standard converters

A range of professional standard converters were considered for this experiment. They can be classified into the two main classes of motion compensated and non-motion compensated systems. Among the tested equipment we can cite the FOR-A FRC 7200, the Snell & Wilcox Alchemist HD, the UKON and the TERRANEX. The test sequences (see description of a set in Fig. 1) were natively generated in 1080i/29.97.

2.2 Test plan

The test plan was designed by an EBU project group to cover several practical conversion scenarios. Two scenarios were considered involving the full supply chain. Potential post-production steps were omitted to clearly identify the impact of the standard converter and its positioning on the final sequences.

The initial test plan considers the US to EU contribution use cases. Only frame rate down-conversion (60–50 Hz) is considered. In future, the inverse path will also be investigated to complete this set of results. The investigated scenarios were the following:

• *Case A.x* – contribution in 720p/50 (Fig. 2): Frame rate and image format conversion happen simultaneously. This case is not currently practically often encountered since most of the HD content exchange is done in 1080i/25. However for the sake of completeness the 720p/50 case is also included.

• *Case B.x* – contribution in 1080i/25 (Fig. 3): Only a frame rate conversion is applied on the contribution feed. This is the most frequently encountered contribution scenario since most of the early HD production was in 1080i/29.97.

For each case above, the position of the standard converter relative to the encoder was considered (before – position 1 – or after contribution codec – position 2).

In the broadcast domain a quality evaluation should always consider the full supply chain. Henceforth, after the contribution simulation, the highest quality feeds were selected and encoded using an H.264/AVC distribution encoder.

2.3 Procedure

The only way to assess standard converter quality performances is by subjective evaluations. Subjective evaluations were conducted at different positions along the supply chain:

1. At the end of the contribution chain (Session 1).

• Comparing outputs of the same converter at the same coding rate (60 Mbps, 30 Mbps) but different positions around the codec (before, after).

2. Stand alone conversion comparison (without any compression) (Session 2).

For the sake of industrial sensitivity the results of this session will not be disclosed.

Figure 1 *Standard converter test sequence samples*

Each sequence above is a mixture of close up, fast panning of the camera together with fast or static objects in the scene

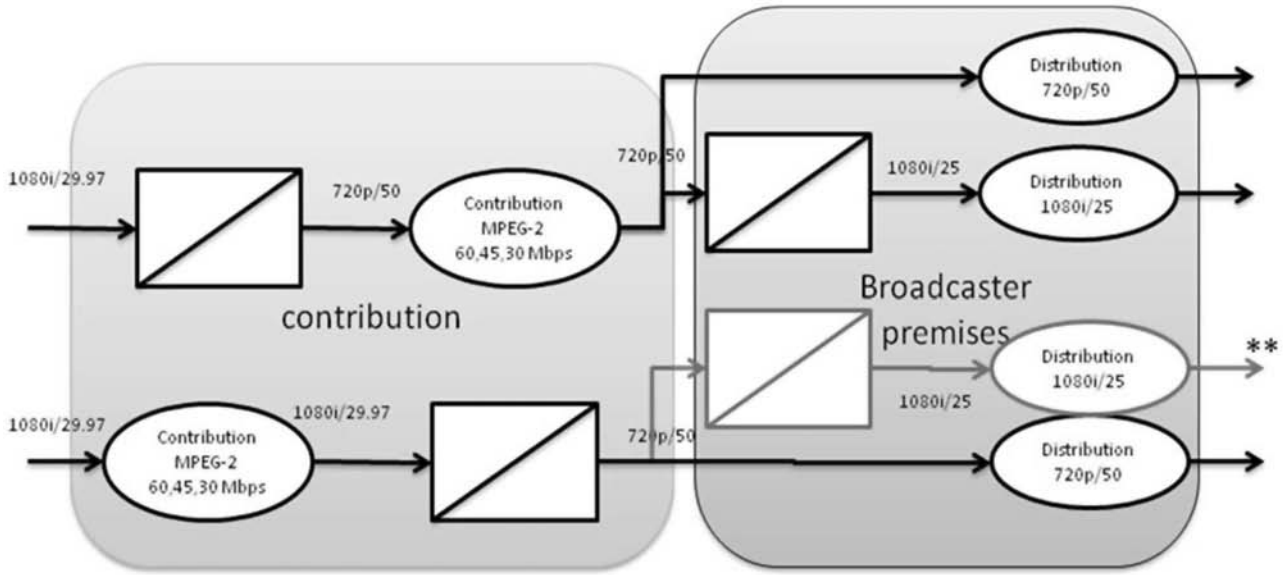

Figure 2 *Case A.x – contribution in 720p50*

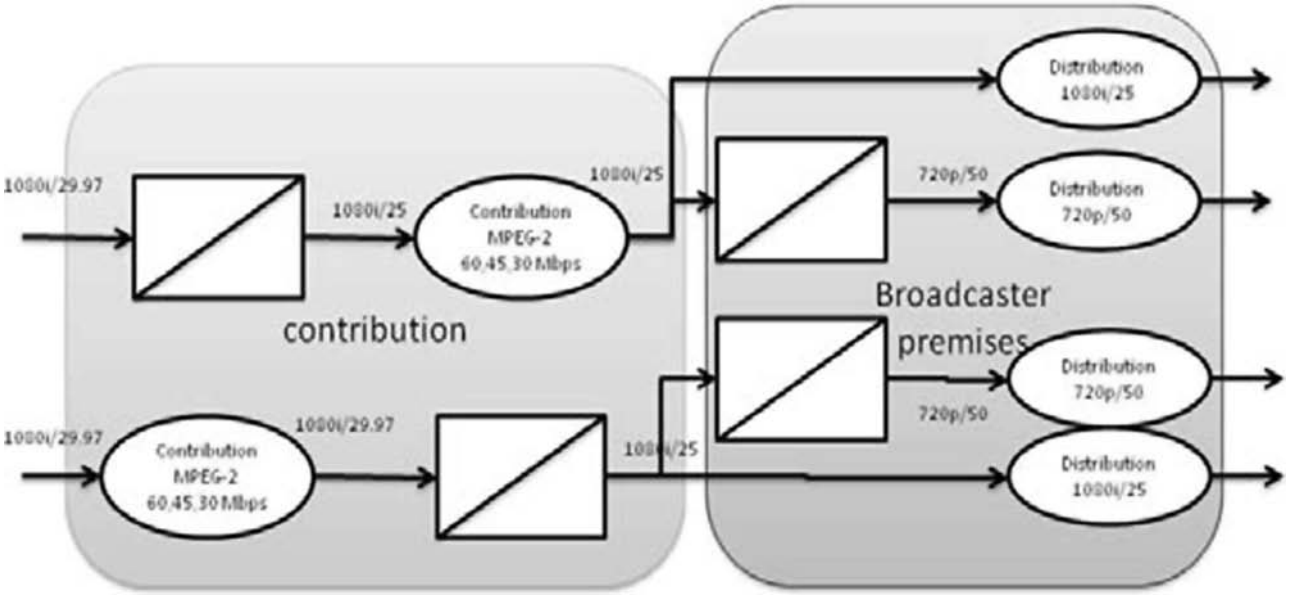

Figure 3 *Case B.x – contribution in 1080i/25*

3. At the end of the distribution chain (Session 3).

- Feeds with only one conversion step.

a. Comparing distribution feeds issued from different contribution paths (converter before or after the encoder) but still using the same converter.

b. Comparing distribution feeds from a contribution feed with conversion before the contribution encoder. The contribution feed comes from different converters output.

- Feeds with two conversion steps (same rate, different converters).

c. Same bit rate, different converters output, contribution feed with conversion before encoder.

The setup for the subjective evaluation involved two 50″ Pioneer plasma displays type PDP 500EX disposed side by side. The pre-recorded sequences were synchronised and played out simultaneously out of DVS Clipster servers. The NTT HE5100 encoder was used for the MPEG 2 encodings.

The contribution feeds were encoded, respectively, at 60 Mbps (Max rate over satellite), 45 and 30 Mbps to assess the cumulative influence of the bit-rate variation with the converter positioning. Only the highest quality contribution feeds were used as source material for the

distribution chain encodings. The latter were done at 12 and 8 Mbps (Video rate) to have again a low- and high-quality end feed. Those bit rates are about the bit rate recommended by EBU for HDTV distribution using H.264/AVC compression (EBU Rec. 124).

2.4 Results

Out of the contribution feeds subjective evaluation, we were able to draw the following conclusion:

• Motion compensated standard converters perform the best frame rate conversion (no or less perceptible image judder). They are thus strongly recommended for video sequences involving high motion such as premium sport events were intensive motion components can be experienced.

• The image judder introduced by the standard converter before the contribution is not masked by the compression at any bit rate. It remains visible (same intensity before or after the codec) throughout the sequence at high or low contribution bit rates. The motion judder introduced by a non-motion compensated system will be interpreted by the compression system as a required motion component. The codec in its attempt to render with the highest fidelity motion, will replicate as close as possible the input judder.

• Having the standard converter before the contribution codec leaves the final contribution feed with just perceptible coding noise and artefacts which can be sometimes interpreted as artificial resolution at higher rates (60 Mbps).

• Having the standard converter after the codec leaves the final contribution feed with less noise but also a just perceptible loss of resolution at higher rates (60 Mbps).

• The two last statements are even more valid as the contribution bit rate rises. The coding artefacts slowly mask the difference in resolution as the bit-rate decreases.

In the second session, the performance of the standard converters was evaluated and compared in a stand-alone manner. Specific behaviours of each system could be identified (loss of resolution, breaking artefacts, blocking artefacts around static object with moving background).

After assessing the influence of the conversion on the distribution feeds (Session 3) it was concluded that:

• The image judder introduced in the contribution domain is not overcome or masked by the compression. It remains perceptible with the same intensity through the whole chain up to the end user.

• Specific standard converter effects can still be distinguished and recognised even after distribution encoding, however this consequence is independent of the standard conversion positioning in the contribution chain. Indeed, the comparison of the distribution feeds at 12 and 8 Mbits for cases A1, A2 and B1, B2 has shown for each of the systems under test that the quality of the feeds was equivalent. However, the specific artefacts of each converter could be identified at almost the same time instant and the same intensity in each feed. Therefore it can be considered that the differences seen at the contribution side with regard to the converter positioning in the chain are not perceptible, after the distribution encoding even at a bit rate of 12 Mbps.

2.5 Future work

• Further comparisons will be conducted in the future to assess the relevance of a 1080i/25 to 720p/50 workflow. It has been demonstrated that in a fully progressive broadcast chain, the 720p/50 distribution format provides up to a 20% bit-rate saving compared to 1080i/25 format distribution. However, some broadcasters plan to produce in 1080i/25 and convert to 720p/50 for distribution, hoping to benefit to some extent from the progressive coding gain. But what will be the actual savings here? Our future evaluation will try to quantify them.

• Since contribution is not only a one-way process, up-conversion (50–60 Hz) may also be investigated in the future for completeness.

2.5.1 HDTV contribution codecs: The HDTV contribution codec topic is being investigated by the EBU N/HDCC work group involving several participating European labs (IRT, RTVE, RAI and EBU). The study in itself is divided into two parts:

• Technology evaluation to provide guidance on the performances compared to existing HD contribution codecs.

• A reference setting test defining the bit rate at which a certain compression system outputs can be considered to fulfil a contribution reference quality that is high quality level providing sufficient headroom for additional post-production. This part involves a cascade with a distribution encoder to assess the quality at the end user side.

In this paper only the first part of the study is described. As stated earlier, it aims at assessing the benefits of new compression technologies such as H.264/AVC, JPEG2000 and Dirac compared to legacy MPEG2 systems. As described below a typical contribution link is simulated by cascaded compression with a spatial pixel shift (scheme described in Fig. 4). The cascade is limited to two generations which is the maximum number of steps possibly encountered on a link.

The test sequences were selected to resemble usual contribution content (sport, live concert etc. – see Fig. 5). They were provided in 10 bit 4 : 2:2 and uncompressed in both 720p/50 and 1080i/25.

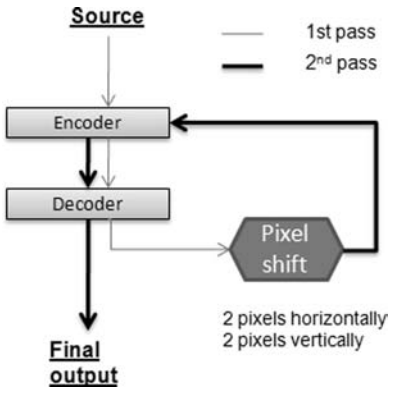

Figure 4 *Test plan*

Some post-production output content was inserted (i.e. the treno sequence). The treno sequence is included as uncompressed, as a fourth generation of AVC-I cascading (100 Mbit/s) and as a fourth generation MPEG2-long GOP (XDCAM HD). Those sequences will allow us to assess the effects of each contribution system on a potential post-production output.

All tested codecs supported the 4:2:2 colour sampling capability which is required for professional contribution application.

2.5.2 Test plan and procedure: Since the aim of the study is to measure the quality gain in using the H.264/AVC or JPEG2000 technologies over the MPEG-2 compression system, we perform encoding assuming in a first step an error-free channel.

According to the codec capabilities, we tested the following bit-rate ranges (15–250 Mbps), knowing that they are in the range of the commonly used bit rates over fibre (>100 Mbps) or satellite (up to 60 Mbps).

Thus we will use different MPEG-2 encodings as comparison anchors: 30, 45, 60 (maximum satellite bit rate on 36 MHz transponder using DVB-S2 Modulation) and 100. Then, we will try to find the equivalent quality using the H.264/AVC starting at half the MPEG2 bit rate and slowly increasing (by steps of 10% of reference MPEG-2 bit rate) it if the quality is not equal or superior to the MPEG2 reference. For example if the reference quality to be matched is a 30 Mbps MPEG 2, then the evaluation will compare it with the following rates 15 Mbps (50%), 18 Mbps (60%) and so on.

The GOP structure tested was the following:

- I – intra-frame only
- IP – low delay mode structure
- IBP
- IBBP
- *Manufacturer:* settings proposed by manufacturers that could provide a good trade off between latency and visual quality (not included in latter results).

GOP alignment was not to be applied during the cascading. The same GOP structure (fixed) is applied to the MPEG2 and the H.264/AVC codec but compared to the I-frame only counterpart on JPEG2000.

2.6 Results

2.6.1 Objective results: As a guideline objective quality assessment metric, we have used the peak signal-to-noise ratio (PSNR). It was used as a trend analyser to have an understanding of the behaviour of the different compression systems under test in a cascading environment.

Figure 5 *Test sequences overview*

Implementation differences: The H.264/AVC implementations considered during this study were objectively speaking, relatively close. The only striking difference was identified for the intra-GOP structure coding were a difference of up to 3.5 dB (3.5 dB 1080i, 2.5 dB 720p50) could be noted (see Figs. 6 and 7). The difference for the over GOP structure was below 0.5 dB which and can be considered negligible. The latter differences where image format agnostic and were still visible at the second generation. This very high degree of similitude re-assures us in the assumption that the result of this study can be applied to other implementations of the same technology.

Concerning their respective cascading losses, a difference of less than 0.3 dB could be identified.

Both systems exhibited a linear, constantly increasing characteristic as shown in the example (Fig. 8).

Concerning the two JPEG 2000 implementations, in addition to the different maximum coding rates, both systems had a different response to cascading depending on the image format under test (Fig. 9). One system had an excellent behaviour using 720p50 (only 0.5 dB lost) while the other implementation had almost the same losses per image format (near 2 dB) (Fig. 10). The latter system also had an unstable characteristic that is the PSNR values were not constantly increasing as the rate was increased. Despite this variability both systems also had a linear constantly increasing characteristic. Only the most stable JPEG2000 implementation was considered during the viewing session.

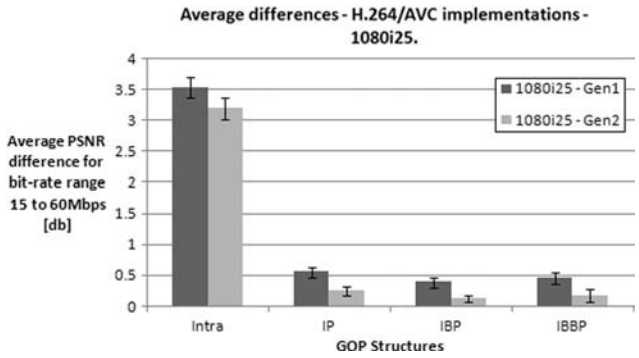

Figure 6 *PSNR difference for H.264 implementation – 720p50*

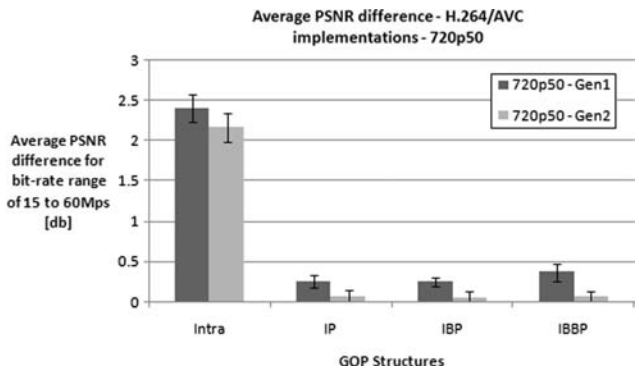

Figure 7 *PSNR difference for H.264 implementation – 1080i25*

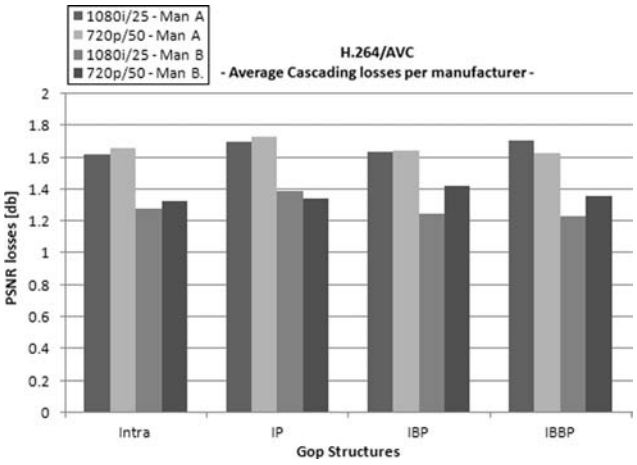

Figure 8 *H.264 average cascading losses*

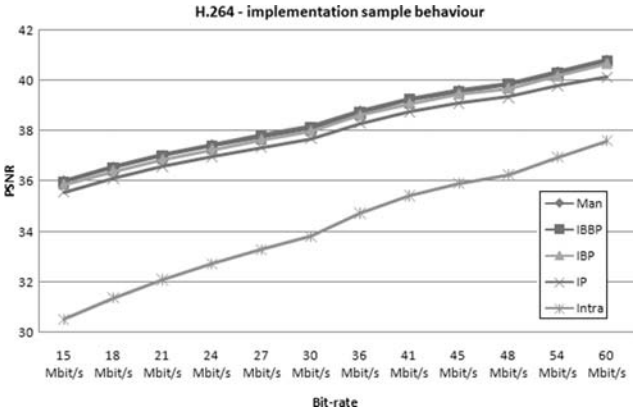

Figure 9 *Example of H.264 system characteristic*

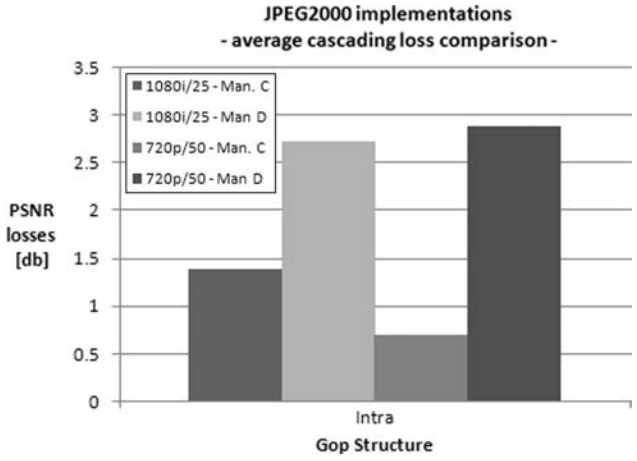

Figure 10 *JPEG2000 cascading losses*

The MPEG-2 system used had similar sustainability to cascading as the H.264 systems (Fig. 11). Therefore the conclusion objectively drawn at the first generation can be applied with confidence to the second generation.

Comparison with MPEG-2: The results of the comparison with MPEG are resumed in Tables 1 and 2.

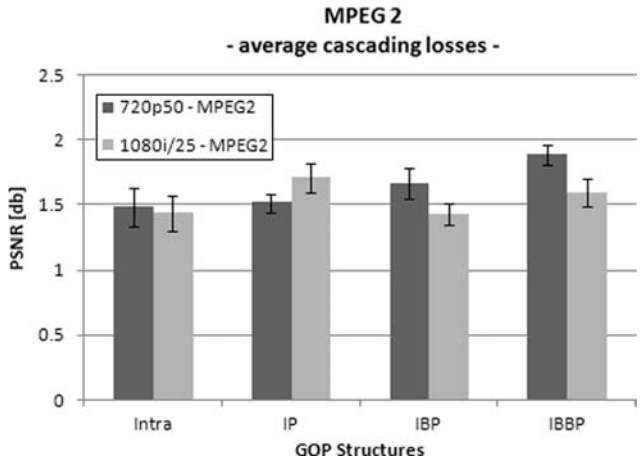

Figure 11 *MPEG-2 average cascading loss*

According to these tables, a minimum gain of 40% can be expected from H.264 with regard to MPEG-2. A decay of 10% can be observed for intra-GOP structures from one generation to the next, whereas it remains constant for all other GOP tested. The gain seems to be even higher for intra-GOP structure and for rates below 45 Mbps. Owing to the drastic difference in coding intra-frames, from one H.264 manufacturer to another this gain should be considered with care.

JPEG2000 shows better gain than H.264 (considering the intra-coding only) on the second generation since it has better sustainability to cascading losses.

Looking at this minimum objective gain, we are tempted to state that the 50% gain of H.264 acknowledged for HD distribution rates can also be met for contribution rates. The subjective evaluation will help clarify this statement.

Subjective quality results: Expert viewings were performed on carefully aligned 50″ pioneer PDP plasma displays using the TSCES assessment method developed at EBU [7] (Fig. 12). The MPEG-2 anchor is shown in the middle screen while both manufacturer implementations are displayed synchronised with the reference at variable rates until a

Table 1 H.264 comparison with MPEG 2

H264	Reference bit rate, Mbits/s	1080i/25				720p/50			
		Equivalence range – Gen1, Mbits/s	Gain in %	Equivalence range – Gen2, Mbits/s	Gain in %	Equivalence range – Gen1, Mbits/s	Gain in %	Equivalence range – Gen2, Mbits/s	Gain in %
Intra	30	<15	>50	<15	>50	<15	>50	<15	>50
	45	[21;24]	<60	[27;30]	<50	[21;24]	<60	[24;27]	<50
	60	[30;36]	<50	[36;41]	<40	[30;36]	<50	[30,36]	<40
	100	>60	N/A	>60	N/A	>60	N/A	>60	N/A
IP	30	[15;18]	<50	[15;18]	<50	[15;18]	<50	[15;18]	<50
	45	[27;30]	<50	[27;30]	<50	[27;30]	<50	[27;30]	<50
	60	[36;41]	<40	[36;41]	<40	[36;41]	<40	[36;41]	<40
	100	>60	N/A	>60	N/A	>60	N/A	>60	N/A
IBP	30	[15;18]	<50	[15;18]	<50	[15;18]	<50	[15;18]	<50
	45	[27;30]	<50	[27;30]	<50	[27;30]	<50	[27;30]	<50
	60	[36;41]	<40	[36;41]	<40	[36;41]	<40	[36;41]	<40
	100	>60	N/A	>60	N/A	>60	N/A	>60	N/A
IBBP	30	[18;21]	<40	[18;21]	<40	[18;21]	<40	[18;21]	<40
	45	[30;36]	<45	[30;36]	<45	[30;36]	<45	[30;36]	<45
	60	[45;48]	<25	[45;48]	<25	[36;41]	<40	[36;41]	<40
	100	>60	N/A	>60	N/A	>60	N/A	>60	N/A

Table 2 JPEG 20000 comparison with MPEG-2 – Intra

J2K	Reference bit rate, Mbits/s	1080i/25				720p/50			
		Equivalence range – Gen1, Mbits/s	Gain in %	Equivalence range – Gen2, Mbits/s	Gain in %	Equivalence range – Gen1, Mbits/s	Gain in %	Equivalence range – Gen2, Mbits/s	Gain in %
Intra	30	<15	>50	<15	>50	<15	>50	<15	>50
	45	[22.5;24]	<60	[22.5;24]	<60	[18;21]	<60	[15;18]	<70
	60	[33;36]	<45	[33;36]	<45	[27;30]	<55	[24;27]	<70
	100	[60;66]	<40	[54;60]	<40	[50;54]	<50	[40.5,42]	<60

Figure 12 *EBU – TSCES assesment stand [7]*

visual match is found by the viewers or cross point in quality can be determined. (System under test becoming better than the reference from one rate to another.)

The subjective evaluation mainly tested the second generation which can actually allow better discrimination of the feeds than the first generation. Only the intra, IP and IBBP GOP structure were reviewed since those are the most relevant structures in contribution. The aim of the viewing was to match the quality of the reference MPEG-2 feed.

The results of the subjective test are shown in Tables 3 and 4. As it can be seen in Tables 3 and 4, the results of the subjective evaluation differ from the assumption drawn by the objective results. The gain varies from one GOP structure to another (intra min 10%, IP 25–30%. IBBP 40–50% at second generation). The gain progression follows the use of B frames in the sequence coding.

H.264/AVC efficiency mainly relies on the use of the toolbox it is provided with. One of the major tools is the hierarchical B frames together with the new macroblock sizes helping H.264/AVC to perform better prediction. Without the predicted frames (B and P) H.264/AVC is merely different from MPEG-2 in intra mode.

Another helper to H.264/AVC gain are the deblocking/denoising loop filters providing H.264 with less annoying artefacts and less noisy sequences than MPEG-2. During the viewing session even if an H.264 coded sequence could have slightly less resolution than the MPEG-2, it was still preferred because it is less noisy. It should be noted that the 720p50 format had better performance than the108i25 by matching the quality equivalence at the lower bound of the defined range. This was not always the case for the 1080i format.

Concerning JPEG2000, the results of the viewing session were drastically different to those of the objective evaluation. This is due to the objective metric used. Indeed, the PSNR is known for not correlating well with subjective evaluation. JPEG2000 process the image by a succession of low- and high-pass filtering providing the output sequence with blurry artefact and cleaner sequences. The JPEG2000 PSNR values can thus appear very high while the structural quality of the video might be different (contouring, overall blur of the sequence with high activity).

JPEG2000 was assessed not equivalent to MPEG-2 quality on specific reference bit rates because it can have different behaviour from one sequence to another. JPEG2000 can have better performance on a set of sequences while being significantly worse on others. For example, JPEG2000 had astonishing performances (even at rate below the reference MPEG-2 rate) on set of sequences such as Olyflags or treno which appeared to be highly critical for H.264/AVC and MPEG-2. Noisier sequences such as crowdrun or parkjoy where more critical to JPEG2000 and drew is performance down. However, as soon as the rate of 60 Mbps was crossed the criticality of the above sequences became less of a problem. This explains the gain of 20% over the 100 Mbps MPEG-2 (for 720p50). The variation of gain compared to 1080i25 might come from the coding method used for the interlaced format.

JPEG2000 implements can be falsely called 'region of interest coding' at lower rates than the reference.

Table 3 JPEG2000 comparison with MPEG-2 – Generation 2

J2K	Reference bit rate, Mbits/s	1080i/25			720p/50		
		Equivalence range, Mbits/s	Gain in %	Comments	Equivalence range, Mbits/s	Gain in %	Comments
intra	30	N/A	N/A	ungradable since reference has too low-quality. JPEG2000 artefacts less annoying	N/A	N/A	same comment
	45	[48;54]	<−20	below 45 Mbps, Better quality for some sequences but worse quality for very noisy sequences such as crowdrun and strong colour bleeding	<50	<−10	equivalence at lower bound
	60	[70]	<−20	below 60 Mbps, Better quality for some sequences but worse quality for very noisy sequences such as crowdrun. Lighter colour bleeding	[60;70]	<−20	
	100	[100;105]	N/A	below 100 Mbps, Better quality for some sequences but worse quality for very noisy sequences such as crowdrun	[80;90]	<20	

JPEG2000 preserved with fidelity the quality of coarse objects in the scenes such as football players while the highly detailed grass was lost. The same observation was made on the crowdrun sequence where the runners were still recognisable. This is simply the graceful degradation feature of JPEG2000 dropping the high-frequency components (highly detailed areas such as grass) of the image. Those details are gradually recovered as we increase JPEG2000's bit rate.

2.7 Conclusion

H.264 can provide bit-rate savings of up to 40–50% depending on the GOP structure (intra min 10%, IP 25–30%, IBBP 40–50% at second generation) at the first and second generation while still providing equivalent quality to MPEG-2.

JPEG2000 needs 20% more bit rate for reference rates below 60 Mbps to provide better quality on MPEG-2 intra. Other metrics should be used to characterise objectively JPEG2000 performances.

The latter results consider the best of the tested implementation and thus a safety margin of 5–10% should be considered to take into account other implementation performances (worse or better).

3 Main conclusion

Standard converters and contribution codecs are often the most important factors influencing the quality of a feed. In our different investigations we were able to determine that the positioning before or after the codec could, respectively, provide a feed with 'artificial' resolution or a feed with less resolution (soften). The difference is more visible at high bit rates than at lower bit rates (30 Mbps) where the coding artefacts may mask the converters effect on the feed.

On a more generic basis, the importance of motion compensation was assessed, especially for events with high motion components.

On the distribution side, the effect of the standard converter positioning in the chain with respect to the codec is barely visible (rather negligible), even at high distribution coding rate (12 Mbps).

The consideration of the 50–60 Hz cases will be conducted later on to complete the test.

As in most media applications, bandwidth is a scarce and expensive resource. The objective and subjective evaluation results have shown that:

Table 4 H.264/AVC comparison with MPEG-2 – Generations

H264	Reference bit rate, Mbits/s	1080i/25			720p/50		
		Equivalence range, Mbits/s	Gain in %	Comments	Equivalence range, Mbits/s	Gain in %	Comments
intra	30	N/A	N/A	ungradable since reference has too low quality. H.264/AVC artefacts less annoying	N/A	N/A	same comment
	45	[41;45]	<10	lower bound with slightly less resolution	[36;41]	<20	equivalence at lower bound
	60	[54;60]	<10	lower bound with slightly less resolution	[54;60]	<10	equivalence at lower bound
	100	>60	N/A	reaching visual transparency	>60	N/A	
IP	30	[21;24]	<30	less noise and better resolution than reference	[21;24]	<30	equivalence at lower bound
	45	[30;36]	<25	less noise and better resolution than reference	[30;36]	<25	equivalence at lower bound
	60	[45;48]	<25	less noise and better resolution than reference	[41–45]	<25	equivalence at lower bound
	100	>60	N/A	similarities on some sequences. Close to visual transparency	>60	N/A	equivalence at lower bound
IBBP	30	[18;21]	<40		[18;21]	<40	equivalence at lower bound
	45	[27;30]	<40		[27;30]	<40	equivalence at lower bound
	60	[30;36]	<50		[30;36]	<50	equivalence at lower bound
	100	>54	N/A	similarities on some sequences. Close to visual transparency		>54	

• H.264/AVC can provide bit-rate savings up to 40–50% over even the second generation MPEG-2 depending on the GOP structure under consideration. The later statement applies for the 1080i/25 and 720p/50 image formats even if the 720p/50 format performs slightly better for the lower bound of the different equivalence ranges defined.

• H.264 has the advantages of providing de-blocking and de-noising filters which provide more benign artefacts compared to the blocky and noisier MPEG-2 outputs. Thus, even if the H.264 had sometimes less resolution it was preferred to the MPEG-2 because of the less annoying artefacts and less noise.

• JPEG2000 has better sustainability to codec cascading. It has even better performances on progressive formats. It requires apparently at least 20% more bit rate to cope with the same quality as MPEG-2 for all sequences encoded. This subjective figure is in total contradiction with the figures drawn from the PSNR curves and emphasises the issue that the latter metric is not that well correlated to the human visual system.

• It must be noted as well that a sequence critical for JPEG2000 might not be critical for H.264 and vice versa. For example, JPEG2000 had a worse result on noisy sequences such as crowd run or parkjoy where H.264/AVC performed relatively well. However on the highly critical olyflags sequence JPEG2000 had excellent performances compared to H.264. Highly grained/noisy sequences can be problematic to JPEG2000.

It should be remembered that even if bandwidth is the most valuable asset on budget calculations for transmissions, other

functionalities and features such as robustness to transmission error, graceful degradation, future proof (scalable) and so forth can be more relevant in applications which are less bandwidth constrained (e.g. regional contribution over radio links or fibre).

4 References

[1] ISO/IEC 14496-10 (MPEG-4 part 10): 'Coding of audio visual objects'

[2] RICHARDSON I.E.: 'H.264 and MPEG-4 video compression' (Wiley & Sons, 2003)

[3] WIEGAND T., SULLIVAN G.: 'Overview of the H.264/AVC video coding standard', *IEEE Trans. Circuits Syst. Video Technol.*, 2003, **13**, (7), pp. 570–576

[4] ISO/IEC 15444-1: 'JPEG2000: image coding system. Core coding system'

[5] TAUBMANN M.M.: 'JPEG2000 – image compression fundamentals and practice' (Springer, 2002)

[6] ISO/IEC 15444-2: 'JPEG2000 – motion JPEG2000'

[7] HOFFMANN H., WOOD D., ITAGAKI T., HINZ T., WIEGAND T.: 'A novel method for picture quality assessment and further studies of HDTV formats', *IEEE Trans. on broadcasting*, 2008, **54**, (1), pp. 1–13

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Decoupling hardware and software in set box top using virtualisation

D. Le Foll

Amino Communications, Cambridge CB24 4UQ, UK
E-mail: dlefoll@aminocom.com

Abstract: In today's modern electronic equipment, the cost of software development for embedded devices frequently represents more than ten times the cost of the development of hardware and associated mechanics. It is also the most common reason for delayed market introduction of products. Decoupling the hardware from the software can produce huge savings, more rapid market entry and a longer product life time. While virtualisation has been used successfully in general IT infrastructure for a few years, its use in embedded software is in its infancy. Initial applications have been targeted principally in mobile and banking applications. Presented is the result of investigatory work carried out to evaluate the value of virtualisation in the development of a digital IP set top box. The goal of this project was to demonstrate the feasibility of running Linux and WinCE in an embedded appliance to enable a maximisation of software reuse and allow legacy software to benefit from new generation hardware. The same principles could be applied to any embedded device running complex software.

1 Introduction

Complexity of software development is growing every day. Embedded devices are now running multi-megabytes of software, which represents tens to hundreds of man years of R&D effort. The difference in complexity between hardware and software tends to make the software the main cause of project delays and failures in most new product developments. This is true even if the software is not visible in the final use of the product (e.g. a TV, car, digital camera, etc.).

The specific difficulty introduced by software is its open complexity. Software is never finished – adding a feature, correcting a malfunction (bug), improving usability – is always desired and can be done. Unfortunately, it is as easy to add a software feature as it is to add a bug, because predicting the consequence of changes in a module can quickly get beyond the reach of the human brain in modern complex software. Proper architecture, modern programming tools, sophisticated operating systems (OS) with memory protection (e.g. Linux) can help to manage this risk, but when a project reaches several tens of millions of lines of code, assuming 100% predictability, there would be a professional fault. This paper will not consider good tools and practices, which still remain the mandatory first step in a proper software project.

In order to mitigate this risk, software engineers have defined solutions that can isolate modules from each other, allowing the decoupling of complex architectures into several smaller subsystems. In this paper, we will look at one of these techniques called hardware Paravirtualisation. Several virtualisation technologies are available and we will explain why this specific one was used for this project.

We will also show how this technology can be used to enable legacy software to take advantage of new hardware capability and extend product life beyond expectation without involving the risk of rewriting the software.

We will focus only on virtualisation models, which have the greatest application to embedded software development.

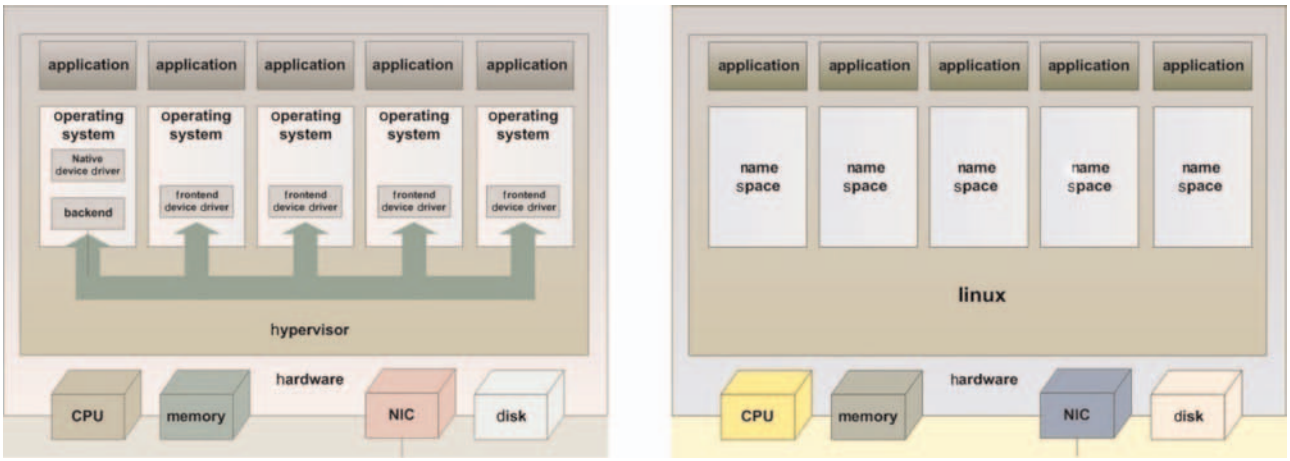

Figure 1 *Virtualisation concepts*

For a more generic introduction to virtualisation refer to the introduction given in Wikipedia [1] (Fig. 1).

2 Various virtualisation technologies

As our goal is to isolate subsystems in a device, we can look at several options ranking from application domain virtualisation down to full virtualisation.

2.1 Virtualisation at application level

Modern electronic devices embed more calculation power and memory capacity than we would have dreamed of having in a research lab in the 1980s. This new reality allows us to use sophisticated OSs based on Unix (mainly Linux and FreeBSD derivatives), which provide full memory protection and shared library services. With these tools it is possible to develop solutions, which can execute several independent instances of shared services.

Good examples of this concept are the concepts of Applets in Java Virtual machines [2] and Virtual hosts in web servers [3]. The isolation provided is limited and typically does not cover the lower layers (closer to the hardware) but can still provide very satisfactory savings. The main advantages of this concept are its simplicity and light weight. The resource sharing and isolation does not increase the memory or CPU usage. Unfortunately, it does need to be incorporated from the outset of the design – which makes this technology of little use when considering the reuse of legacy software.

2.2 Virtualisation via name spaces domain

Unix OSs and in particular Linux, which is very popular in embedded software development, offer the possibility of isolating programs between themselves via the concept of name space. In a nutshell, if two software processes are created in two different name spaces, they do not see each other. The solution was originally developed for security reasons, but can be easily redeployed to isolate subsystems in order to provide greater isolation between subsystems. This model of virtualisation is extremely light weight, as most of the OS is shared between the domains. It also provides the option, depending on its configuration, of fully isolating subsystems from each other. Its main limitation is the requirement for all applications to share the same version of the OS, which means that it is necessary to either upgrade legacy codes to run on newer versions of the OS or to back port new applications to an older version of the OS.

The project vserver provides an implementation of this concept that would gain from being known in the embedded developer community [4].

2.3 Virtualisation via the OS

By promoting the concept of Name Spaces Domain and using the memory management unit (MMU) in a very smart way, developers have created an option that allows running Linux under a Linux host. This is the project Open Virtual Zone (http://en.wikipedia.org/wiki/OpenVZ). This model allows the running of different versions of Linux on the same host without the full duplication of the resources. While this virtualisation solution is very efficient for the CPU it does require a duplication of the RAM for many parts of the OS.

2.4 Full virtualisation

As modern CPUs provide the possibility of trapping any illegal memory and IO access we can also run several client OSs on the same CPU. This works as long as during the initialisation great care has been taken to declare any access to the critical hardware (e.g. the MMU, Hard drive, ...) and to reserve it exclusively for the virtualisation hypervisor (host). When a client OS tries to access a critical section, its call is trapped by the CPU and rerouted to the host, which will execute the requested service (e.g. write a file) and present the result of the trapped instruction exactly as

if it had been executed directly. This model allows the running of client OSs, which have no knowledge of the fact that they have been virtualised. The commercial product Vmware [5] and its open source competitor Virtual Box [6] are two well known implementations of this concept.

In this model, no part of the client OS and associated application are shared and the system must provide enough RAM to support this duplication. On the opposite side the dependency between the client OSs and their associated application is nil, which make the full virtualisation a great model to support legacy applications. Furthermore, to ease the installation and provide the base services required by the client OS (file and hard-drive management, GUI, keyboard, networks, etc.), the host is also a full OS (Linux, Solaris, Windows).

The power of this model is that it delivers a complete independence between the hardware and the host client OS, which does not have to be modified to be executed. It comes at the cost of performance, as the redirection of each critical routine has to be done to keep this full transparency. Some of the latest generations of CPUs (Intel, AMD) have implemented special instructions, which can improve the performance of this model. This solution is still generally too heavy to be used in embedded systems but could start to be applicable in the years to come.

An extension of this model called Emulation, where the host emulates the CPU model by software required by the client, can be used to support client OSs, which are incompatible with the host CPU. This model has a huge CPU overhead and is only interesting for supporting legacy codes which cannot be recompiled. It is generally not applicable in the embedded domain (Fig. 2).

2.5 Paravirtualisation

The concept of Paravirtualisation simply reuses the full virtualisation model but modifies the critical section of the client OS to call the emulated function from the host, in lieu of using the redirection via a CPU Illegal Instruction Trap (http://en.wikipedia.org/wiki/Paravirtualisation). As the client OS must be modified, Paravirtualisation can only be used for client OSs that are delivered with the source code for the critical sections. Fortunately, in the embedded domain, this is generally the case with open source OS (e.g. Linux) and with commercial software (e.g. WinCE, WxWorks). Initially Paravirtualisation was developed to offer better performance than full virtualisation for web hosting, and the open source project Xen [7] supports several large server deployments over the world.

Within this limit, Paravirtualisation provides the same isolation as a full virtualisation but with far smaller CPU overhead (RAM overhead remains comparable) and for this reason is far more applicable to the embedded domain. Furthermore, some smart people realised that in the embedded domain you might not need a host that provides a large number of base services. Consequently, it could be valuable to enable sharing of the hardware by several client OSs without the need to provide any shared high level services. Thus hardware Paravirtualisation was born (Fig. 3).

Hardware Paravirtualisation does offer a full independence between the client OS and its associated applications at a very

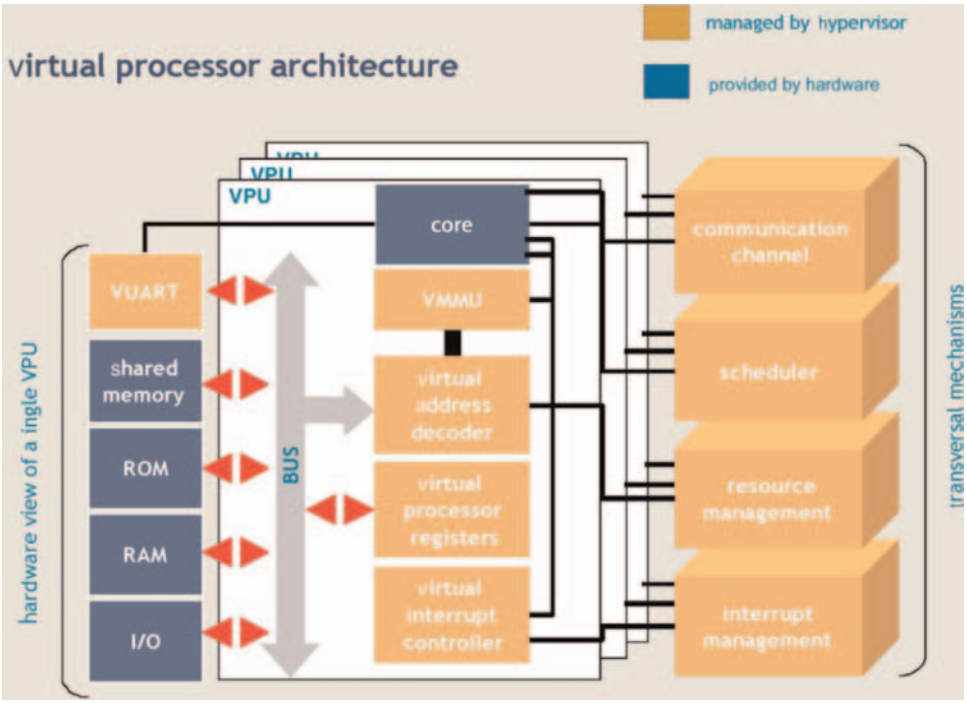

Figure 2 *Virtualisation architecture*

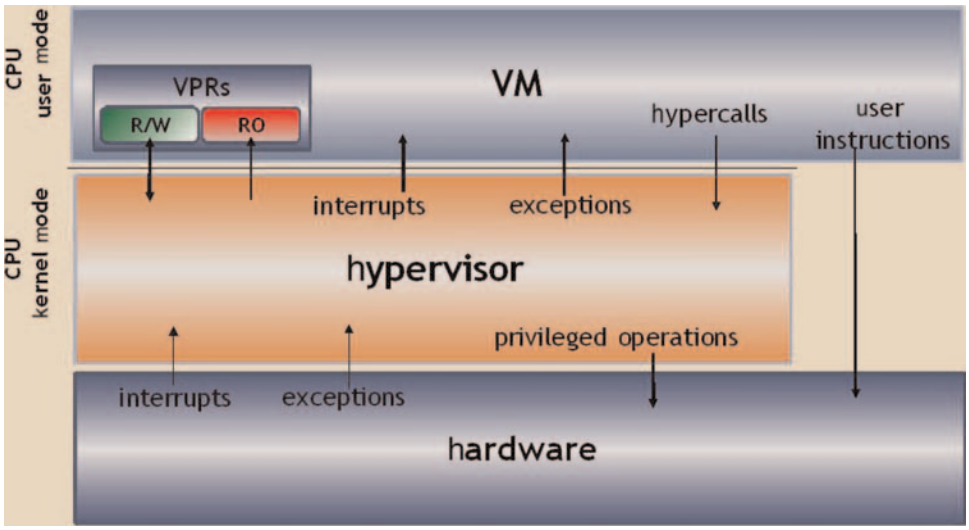

Figure 3 *Paravirtualisation*

low CPU overhead. The RAM footprint is mostly limited to the client OS, because the host OS does not exist and is replaced by a very light hypervisor module.

This hardware Paravirtualisation is the solution that we will describe later in his paper, because it is the only one today that is really applicable in the embedded domain.

3 Hardware paravirtualisation

3.1 Set top box context

A set top box (STB) is a device dedicated to playing video and audio signals received from either a broadcast signal or the Internet network (IP) on a television. It is normally built around a dedicated CPU called a system on chip (SoC) which offers a Mobile phone class CPU engine associated with dedicated hardware helpers for video decode and signal deciphering. The OS is nowadays mostly Linux, but Real-Time OS (RTOS) and WinCE can also be seen. The RAM capacity varies from a few tens of megabytes for STB based on RTOS, to a few hundreds of megabytes for the Linux implementations.

The complexity of the software in STB can easily represent several hundreds of man years of effort and the very tight resources imposed by the low price of these devices, often forces engineers to implement optimisations that are very specific to their SoC model.

3.2 Virtualisation use case for an STB

The diverse objectives of the creating of new products supporting new features – while keeping legacy software untouched for older feature support and allowing the use of latest generation of SoC – provide new opportunities. Producing a technical solution to meet these objectives can allow more rapid, lower cost market entry for new products (Fig. 4).

In this ideal world, older software would become high margin revenue makers instead of the situation we have today where it is necessary to maintain and port to new architecture (SoC and OS) killing margins because of the associated high maintenance costs.

In our project, we decided to validate the feasibility by implementing WinCE5-based STB software on a modern multi-core SoC (BCM7504) using Linux to provide new services that were not supported by our legacy software.

The WinCE5 legacy had to support basic video display and some interactive applications such as an EPG, while Linux would offer DLNA connectivity and TR069 remote maintenance.

As WinCE5 does not support multi-core architectures and supports only FAT file systems, we wanted to offset some of the CPU intensive tasks from WinCE5 to Linux and implement a server class file system (JFS).

Because STBs are constrained by the new European Code of Practice on low energy, we wanted to be able to shut down one of the CPU cores during low activity periods in the most transparent way possible.

3.3 Our feasibility implementation

After careful study of the various virtualisation alternatives we selected a hardware Paravirtualisation solution offered by Trango-vp which is now part the VmWare group [8]. As we had access to the source code of both target OSs (Linux and WinCE5), hardware Paravirtualisation gave us full independence for our software without taking too much CPU power (<3%). We had to provide enough RAM to load both OS and associated applications. The cost of 512 MB RAM was considered very small in comparison to the potential saving on software engineering.

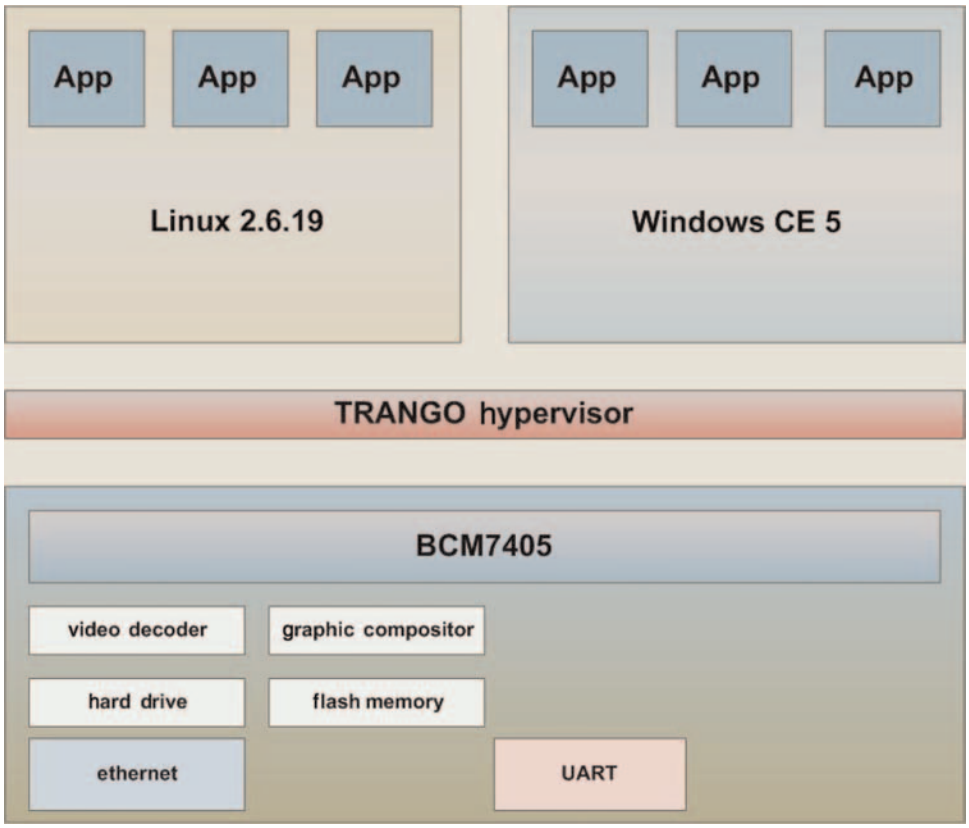

Figure 4 *STB use case*

We decided to dedicate the control of the video rendering, network access and file system to Linux and let WinCE5 access these services via specialised drivers. This model is viewed by WinCE as if hardware helpers were present in the system. This allowed Linux to benefit from the multi core CPU to maximise the use of the resources offered by this new SoC generation.

3.4 Trango-vp Paravirtualisation concepts

Trango-vp hypervisor is a small software module of just a few tens of kB, which provides the following management services:

- memory
- interruptions
- virtual CPU
- virtual CPU cache
- shared memory
- communication link
- priority mapping
- profiling
- dynamic resources.

Each client OS has to be modified to call the hypervisor in related critical sections. As the number of supported services is limited, these modifications are quite light compared with a port for a server type Paravirtualisation of the type you would find with Xen.

Drivers must be written to redirect selected WinCE5 services to Linux.

3.5 Findings

To the surprise of many sceptics, the solution delivered what had been expected. We could demonstrate the simultaneous decode of a high definition video streamed from an Ethernet interface while running two graphics applications. Half the work being done under WinCE5 and half under Linux.

We proved that the WinCE and Linux applications would run without modification in total ignorance of the underlying virtualisation.

The CPU overhead was negligible and the RAM usage is limited to the need of each client OS. The RAM and CPU overhead of the hypervisor is negligible.

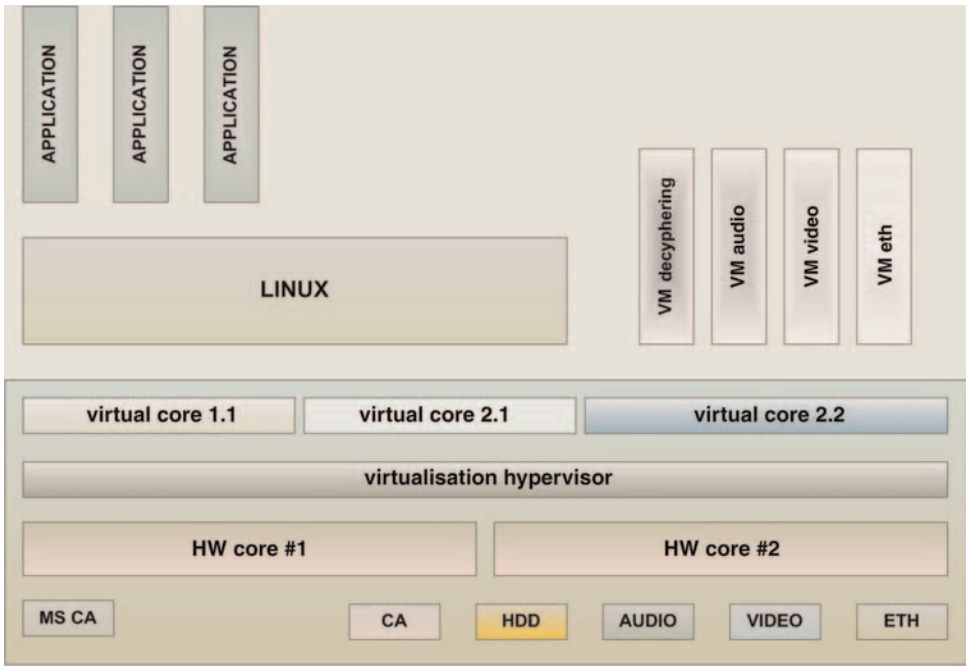

Figure 5 *Virtual hardware helper*

3.6 Applicability

Reducing the cost of legacy maintenance is not an option, it must be done. Virtualisation is the tool with the muscle to achieve this objective. It is the only solution in case of legacy software running on two different OSs like WinCE and Linux. When all applications are Linux based, then virtualisation at application level or server level is an alternative that it is worthwhile considering.

Hardware Paravirtualisation can also open access to new SoC facilities that cannot be supported by legacy OSs (e.g. WinCE5, multi-core CPU support). For these situations, virtualisation can boost the performance of a solution at a greatly reduced cost and is an option that cannot be ignored.

Hardware Paravirtualisation can also be used to offset the implementation of low level real time software under Linux. Development of critical sections must be done in the Kernel domain where traditionally tools offer limited functionality and where skilled engineers are very hard to find.

Paravirtualisation allows you to create a dedicated client, which is purely based on interruption, and can offer a virtual hardware helper to Linux. This architecture could save several hundreds of software man years of embedded Linux development. The complexity of playing in Kernel mode is generally under-estimated and is usually paid for at maximum cost during the later stages of the project (Fig. 5).

3.7 Limitations

The only real constraint imposed by hardware Paravirtualisation is the extra memory requirement. The port of the hypervisor and the adaptation of the client OS can be done by the technology partner. As the customisation of the client OS is quite light, these costs remain acceptable in relation to the performance delivered by the solution.

The sophistication of the technology can be frightening at a first examination, but a simple check on the cost of legacy software maintenance and uplift, should motivate even the most conservative management team.

If you wish to learn more about this topic, read the hardware Paravirtualisation FAQ at [9].

4 Acknowledgments

I wish to thank Jean François Roy Application Engineer of Trango-vp (Fr), Alexandru Faliboga SW Architect at Deuromedia (Ro) and Ly B. Tran VP of Engineering at Broadcom (USA) without whom this investigation could not have been done.

5 References

[1] http://en.wikipedia.org/wiki/Virtualization

[2] http://java.sun.com/applets

[3] http://httpd.apache.org/docs/2.2/en/vhosts/examples.html

[4] http://wiki.linux-vserver.org/Paper

[5] http://en.wikipedia.org/wiki/Vmware

[6] http://en.wikipedia.org/wiki/Virtual_box

[7] http://en.wikipedia.org/wiki/Xen

[8] http://www.vmware.com/technology/mobile/index.html

[9] http://www.fridu.org/dominig-posts/48-virtualisation embedded/101-embeddedvirtualisationfaq

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Novel approach to forensic video marking

N. Thorwirth

Verimatrix, Inc., 92121, USA
E-mail: nthorwirth@verimatrix.com

Abstract: Digital watermarks provide a way to bind a video copy to a recipient, allowing for tracking of illegal distribution and creating an effective deterrent against piracy. This encourages responsible consumer behaviour, rather than restricting the consumer's media use. These types of techniques can be seen as a replacement to traditional DRM that restricts users or as an addition, that allows tracking of the content after it has been decrypted and re-recorded, for example, through the 'analogue hole', when camcording the TV screen. For practical deployment, the marking needs to be invisible, secure against removal and robust against a large variety of distortions, and it must be applicable in existing consumer devices. This study outlines research results that demonstrate efficient embedding of information that can be restored to a human-readable form, which is much more powerful than recognition by a computer, creating a mark that remains identifiable after distortions, such as camcorder capture and compression. Several concepts are introduced that have been combined to overcome the challenges of making this video marking technology robust, secure and yet invisible.

1 Introduction

The use of digital video has grown massively in diversity as well as popularity and is now approaching the level of digital music in its ease of use as well as its ubiquity. The advantages of digital distribution, just as with music, fuel the growth of large-scale piracy. The rightful content owners are missing out on revenues and could be discouraged from future investments in the creation of content. Digital video delivered to a set-top-box (STB) provides a unique environment for content protection using user-specific marking, potentially providing a superior distribution platform that is preferred by movie studios that are looking to secure their sensitive early release windows. Early release windows translate into more demand for pay-TV operators and they are a crucial component to increase the momentum and customer base of this distribution channel.

User-specific digital watermarking is ideally suited for forensic applications that aim to identify the source of piracy. It acts as a deterrent because it registers the individual video to its owner. The embedding task is performed in each STB or DVR, allowing individually marked copies, ideally without the requirement to process individual videos at the head-end for each receiver. The extraction can be performed on any copy that is in public distribution and originates from a protected piracy source. To identify the embedded information, the copies are read in a central location avoiding the distribution of the extraction application. This prevents a possible attacker from verifying the success of modification applied to the video with the intention to remove the mark. Technologies are required that enable content marking in a secure, robust and unobtrusive manner.

2 Challenges of content tracking technologies

Several approaches have been applied to secure digital media: digital encryption technology is effective to enable secure delivery. However, once decrypted and presented in a human visible format, it can be re-recorded to obtain and distribute an illegal, unsecured copy. No effective technology exists today that prevents re-recording using a camcorder.

Unlike encryption, the marking of media does not prevent distribution but allows investigation to identify individuals who are responsible for illegal content use, by embedding recipient information in the media.

One way of marking media is by including information in the media file that is ignored during playback, but can be extracted from the original digital file. Apple iTunes is widely assumed to be using this method to tag DRM-free file distribution [1]. These tags enable identification of the unmodified file, but are easily removed and they are destroyed when the file is re-recorded or converted to another format.

For more robust marking, visible text or dots have been used to carry identifying information in video. Although this information does survive re-recording, it can easily be identified and removed in order to disable the ability to track the content. As these marks are observable and not part of the original content, they also degrade the consumer experience.

Digital watermarking is another marking approach that has been suggested in many variations. Common digital watermarking schemes embed information by introducing manipulations at defined locations in space or time. These watermarks are detected by watermark detection software [2]. In other words, it is a process that modifies media content by embedding an invisible machine-readable code into the actual content.

When interpreting the manipulations during the extraction of the information, some knowledge about the insertion positions is typically required. When the positions are modified, that is, misplaced or weakened, the readout becomes unreliable or impossible. Such modifications do routinely occur during simple media treatment in preparation for illegal distribution: cropping, change of aspect ratio, frame rate change and conversion to another file format, including lossy compression, during which perceptually insignificant information is eliminated to reduce the size of a digital media file. Camcorder capture from a display device combines these transformations with additional degradations. Relative misplacement of the mark and underlying content can also be created intentionally by imperceptible, slight or combined modifications, such as shifts, rotations and time jitter. Publicly available tools [3, 4] apply these modifications, as benchmarks or attacks in an automated fashion. Since current image processing algorithms are not very successful in recognising misplacements in distorted content (a process also called registration), these modifications render these types of machine-readable digital watermarks ineffective.

Digital still images have been the early focus of watermarking research and commercial exploitation [5]. Video watermark approaches have often been based on the application of the image watermark applied to video frames. Although this is the natural progression and allows the embedding of large amounts of data, this approach does not efficiently use the time domain for gathering embedded information because detection is only successful if some information in individual frames can be recovered. The approach fails when none of the watermarks can be read at least in part because of a failure in registration or destruction of relevant areas.

Digital watermarking typically involves a complex transformation of the original image and of the message to be embedded in order to allow invisible embedding. These transformations often challenge the processing capabilities available in consumer devices that have not been designed to include this functionality.

For recognition of the mark, the transformations are inverted to extract the embedded information. This approach enables important clues for an attacker to find a weak spot in the extraction when analysing the embedding system. Ideally, the extraction function is independent of the embedding; leaving the attacker no clues as to how the extraction application is applied. This aspect is particularly relevant for watermarking that is deployed in an environment where either the embedder or the detector is included in consumer devices and may be subject to reverse engineering.

3 Three elements of the solution

To overcome the challenges outlined above, a new approach is required that takes the specific requirements of forensic watermarking of video content into account.

3.1 Human readout

The approach starts with the realisation that the recognition of distorted content is a task that humans can still perform better than machines. This fact is used for example in the so-called CAPTCHA images [6] (completely automated public turing test to tell computers and humans apart) that blocks website robots and identifies human users. This is illustrated in Figs. 1 and 2 that show the difference between machine-readable information, such as a barcodes and human-readable text. While the barcode becomes unintelligible after a ripple transformation and an overlaid

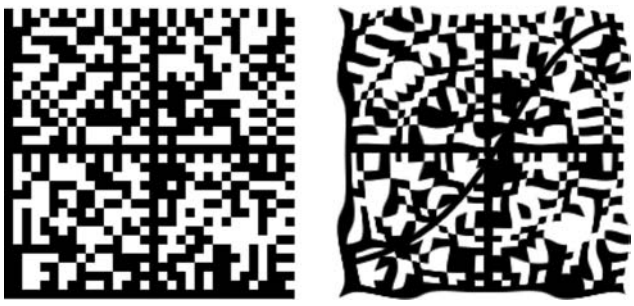

Figure 1 *Distortion prevents readout*

Figure 2 *Information remains human-readable*

line, the text in Fig. 2 can still be read after those distortions. So instead of embedding machine-readable information, a human-recognisable image is invisibly embedded in the video and distributed over time. During extraction, that information is aggregated in order to retrieve the human-readable image containing the embedded information.

During extraction, the marked video is processed by a computer, using an image process that merely emphasises the embedded information but does not aim to detect it. The actual interpretation of the mark is performed by a human, making it possible to detect the mark in degraded content without explicit knowledge of transformations that have been applied to the original content. Although it is very difficult for machine-readable watermarking technology to interpret a misplaced mark, a message can be easily read by a human even if it is changed in size, bent, stretched and on a noisy background. Any distortion that will maintain a reasonable quality of the video and does not transform it beyond recognition to a human observer will allow human interpretation of the embedded message. The shape of the video and that of the mark are coupled in a way that does not allow destroying one without the other.

3.2 Disassembling the mark

The second element of the described solution is to spread the mark over time and to embed different portions in each frame in a way that the mark can be re-assembled over time. The mark cannot be recovered from a single frame, but each frame contributes to the detection result, which is an accumulation of all video data. This distribution is applied for the purpose of security, in order to be robust against attacks that aim to determine and distort embedding locations for each frame.

A related concept of securing information is called visual cryptography [7]. Two images that have perfect random properties will reveal a message when overlaid. Figs. 3 and 4 are examples for two images that will reveal a message (Fig. 5) when aligned and overlaid, such that the white areas are transparent. In locations of the background, the two visual keys consist of an identical pattern. At the areas of the foreground the information in the keys is inverted. The combination created a black foreground on a noisy background, allowing readout of the information.

The distribution of the embedding information is random and only used during embedding. It is not required for the readout of the mark, thereby preventing a deterministic process that could serve as a point of attack in an attempt to determine and distort the marking locations.

Figure 3 *Visual key 1*

Figure 4 *Visual key 2*

When combining this concept with the marking of video, the modifications are embedded in the video where the actual video content is the dominating component of the visual data. The mark represents a portion of the frame content that is too small to be observable, which prevents analysis on individual frames and determination of the random locations. As a result, the mark is securely hidden and protected against deterministic removal of the mark because of non-deterministic embedding variations and is also protected against analysis of a frame because of its subtle nature.

3.3 Aggregation over time

When uncorrelated images are blended together, the resulting average image will increasingly resemble a flat grey image, with pixel values close to the mean, as shown in Fig. 6, depicting 12 images above a blended version of those images. Although the frames of a movie are not all uncorrelated, they do represent a large number of images. This effect is observable in particular in movies that provide a high degree of variation within and in between scenes. The information embedded in the content in contrast is at constant locations and therefore will increase in relative signal strength as the impact of every single frame diminishes. A simple averaging though is not sufficient to expose a readable mark that has been embedded in an invisible level. To actually archive robust detection, additional filtering is required that enhances the signal of the embedded mark in every frame before the frames are merged. The filter reduces the contribution of the underlying content on the combined image while at the same time improving the effect of the mark in individual frames. Although several filtering systems are suited to emphasise the mark and they are most effectively applied in concert, the fundamental concept is to highlight high frequencies in every frame to emphasise the slight variations added by the mark over the actual content that is dominating with large, low-frequency content. This filter also contributes to the faster convergence of the combined

Figure 5 *Combined visual keys 1 and 2*

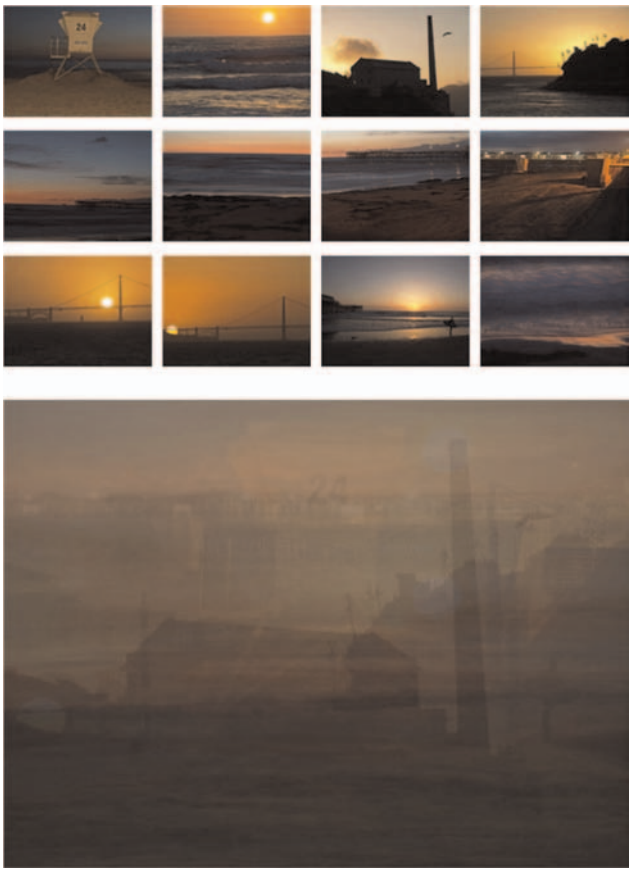

Figure 6 *Image blending*

image by emphasising fine details that are typically subject to more frequent variations than larger content elements.

4 Invisible embedding

In addition to the selection of embedding locations for security as described above, the embedding locations are also varied according to the estimated visibility. This estimation is performed by the perceptual model. The perceptual model identifies locations where modifications will stay below the just noticeable threshold and thereby make the modification invisible to the consumer.

A fundamental concept for the design of the perceptual model was the fact that noise is less noticeable in the presence of already existing noise. More generally expressed by the Weber–Fechner law that states that a weak signal is less perceptible in the presence of a strong signal.

For the purpose of establishing the perceptibility of the mark, the present noise is measured for all pixels of a frame and the allowed modification is determined accordingly. Spatial noise is measured as the variation of proximate pixels within a frame and temporal noise is measured as a variation of pixels in the same location at different times.

Another element that determines invisibility of colour changes is the distance between neighbouring colours as represented on the display device. The perceived difference between neighbouring points in the colour space varies depending on the display technology such as a CRT display, plasma or projection. If the display device is known, the perceptual model takes these differences into account and adjusts the amount of embedded modifications accordingly.

These three elements are combined and weighted in order to determine the allowed modification for each pixel in a frame. Details on how to measure each noise level and the combination have been determined by comprehensive subjective testing and invisibility evaluation of different content in different viewing conditions and display devices. The result is not an exact science with invisibility results that can be objectively measured; however, it is tuned and confirmed with the help of expert viewers that are trained to detect content variations and artefacts.

5 Results

The resulting solution shows remarkable robustness, even in the presence of modifications that would suggest a removal of the minute, unnoticeable marking additions. These degradations include compression of the colour domain or analogue transformation. The reason for the approach being effective in these scenarios is that a fraction of the modification will survive any transformation that leaves the content in a reasonable quality.

The specifics of the technology do not allow for automatic and objective measurement of the actual extracted signal strength as it is part of the design that the information is read by a human. Although the mark can be measured for verification by correlation to the known embedded information or OCR can be applied for reading, this will not be adequate to determine the level of human readability. Multiple third-party tests have confirmed robustness of the marked content against attacks of camcording, variations of the shape and size of marked content and compression, to levels used on peer-to-peer sites. The time it takes to extract the mark varies with content degradation and it is remarkable that the extracted mark continues to improve even after several minutes and thousands of digital frames of content have been aggregated during extraction.

The security of the extraction is established because of its asymmetric nature. First, the readout is done by a human and the embedding by a machine. Secondly, the extraction filter is mostly independent of the embedding and does not follow the specifics used during embedding. For instance, the embedding locations are spread in a pseudo-random fashion that is not known during extraction. Both levels of asymmetric embedding and reading increase the security against attacks that aim to understand and to invert the embedding algorithm. It also provides robustness against the so-called oracle attack that degrades the watermarked

content in an automated loop to find the least amount of degradation that causes the detector to fail to recognise the mark.

Unlike machine-readable digital watermarking, the detection process displays an obvious and unambiguous human understandable outcome, for example, a serial number. When uncovering the mark, the embedded graphic slowly appears from the marked video, showing that the mark is derived from the content. The extraction can be interpreted or read by a layman. The result is easy to understand and can be used as persuasive evidence of wrongdoing during an investigation into content misuse.

6 Conclusion

We have shown how to combine several elements that are novel to watermarking in a solution in order to provide the level of security and robustness that is required for a forensic, user-specific marking application. Security is accomplished with random variations and an asymmetric process. Robustness is accomplished through accumulation of information over time and enabling human readout. Although this key element of human interpretation seems counter intuitive in an environment that aims to be increasingly automated, it overcomes the otherwise significant challenges of content misalignment that fools machine-readable approaches. The brief human interaction that is required to interpret the mark is not a significant hurdle in the forensic application where all content is marked but only illegal distributions are examined.

With the obvious need for better protection against movie piracy, robust and secure marking is an effective deterrent that encourages responsible consumer behaviour and alleviates the revenue loss that content owners are currently facing. The way content is used is not restricted to a-priory implemented use cases, yet watermarking may be an additional important factor for the bottom line – maintaining the consumer motivation to actually *pay*-per-view for premium content instead of trawling for 'free' peer-to-peer downloads. Effective content protection can make the difference between a few subscribers that distribute movies to a large Internet community and a growing consumer base that values premium content.

7 References

[1] ORION E.: 01/2009, DRM-free iTunes files have your number. Available at: http://www.theinquirer.net/inquirer/news/357/1050357/drm-free-itunes-files-have-your-number, retrieved on 1/29/09

[2] ZHAO J.: 'Look, it's not there', *BYTE Magazine*, January 1997

[3] FABIEN A., PETITCOLAS P.: 'Watermarking schemes evaluation', *IEEE Signal Process.*, 2000, **17**, (5), pp. 58–64

[4] PETITCOLAS F.A.P., ANDERSON R.J., KUHN M.G.: 'Attacks on copyright marking systems'. Proc. Information Hiding, Second Int. Workshop, IH'98, 1998 (*LNCS*, **1525**)

[5] MOHANTY S.P.: 'Digital watermarking: a tutorial review'. Department of Computer Science and Engineering, University of South Florida, Tampa, 1999

[6] VON AHN L., BLUM M., LANGFORD J.: 'Telling humans and computers apart automatically', *Commun. ACM*, 2004, **47**, pp. 56–60

[7] NAOR M., SHAMIR A.: 'Visual cryptography'. Advances in Cryptology – Eurocrypt 94, 1994 (*LNCS*, **950**), pp. 1–12

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronic Letters*

ISSN 2041-5923

Near-live football in the virtual world

M.J. Williams[1] *D. Long*[2]

[1]*Sony Broadcast and Professional Research Labs, UK*
[2]*Sky Sports, BSkyB, UK*
E-mail: mike.williams@eu.sony.com

Abstract: Live television coverage of football matches has now reached a sophisticated level with multiple manned and unmanned cameras positioned at strategic positions around the stadium. High-definition (HD) cameras and a HD live production chain ensure excellent picture quality for the viewer in the home. All significant events during the match, such as goals, are captured by many cameras allowing a choice of replays from several fixed camera positions. Described is work that is being undertaken to create new virtual camera positions anywhere around or within the football pitch. For example, in some situations it would be desirable to position the camera behind a linesman to validate an off-side decision or behind a referee to validate a penalty decision. The approach that has been taken in this work is to accurately extract the co-ordinate positions of the players, officials and the ball on a frame-by-frame basis and use this information to configure a virtual world where a camera can then be arbitrarily positioned. This football metadata has to be produced with low latency and with a high accuracy to faithfully represent the play in the real world. The virtual world itself must be compelling by providing photo-realistic representations of the players and stadium as well as realistic player animation characteristics during normal play. Described will be a project partnership between Sky Sports and Sony. Details are provided about the camera acquisition and calibration system, the image processing required for player and ball tracking as well as the event annotation that all together encapsulate the football metadata to describe an event such as a goal. The mechanism to use this metadata to faithfully represent real play in a football rendering engine such as a computer gaming engine with a short processing delay will be described.

1 Introduction

Sports and data have always gone hand in hand whether you are a supporter in a stadium or watching it in high definition (HD) in the comfort of your own home. With the advent of sports production in HD, the industry is now looking for ways to enhance their data gathering capabilities. The ability to create football metadata at the level of fidelity to allow a faithful reconstruction of a football match in a high-quality computer gaming engine not only has the potential to enhance the viewer's experience but also has the potential to merge live sports with online communities and video game fans.

The vision that drives this project is to create a photo realistic, high realism virtual football match in near real-time to enable the coexistence and fusion of the real world with the virtual world.

Presented here is a sports graphics system capable of producing a computer generated virtual replay of chosen events that occur during a football match. One of the key features of the described system is the ability to seamlessly mix from a real camera view to a computer generated virtual view while the video is playing. This provides a point of context for the viewer enabling them to be convinced by the transition from real to virtual and back again (see Fig. 1).

Based on image processing technologies, the system employs an HD camera system that captures the entire field of play for the duration of the match. Along with this video the positional data of players and ball is generated and stored to allow access to any event involving any player throughout the match. Potentially, any moment of any match during a football season can instantly be recalled for review or analysis.

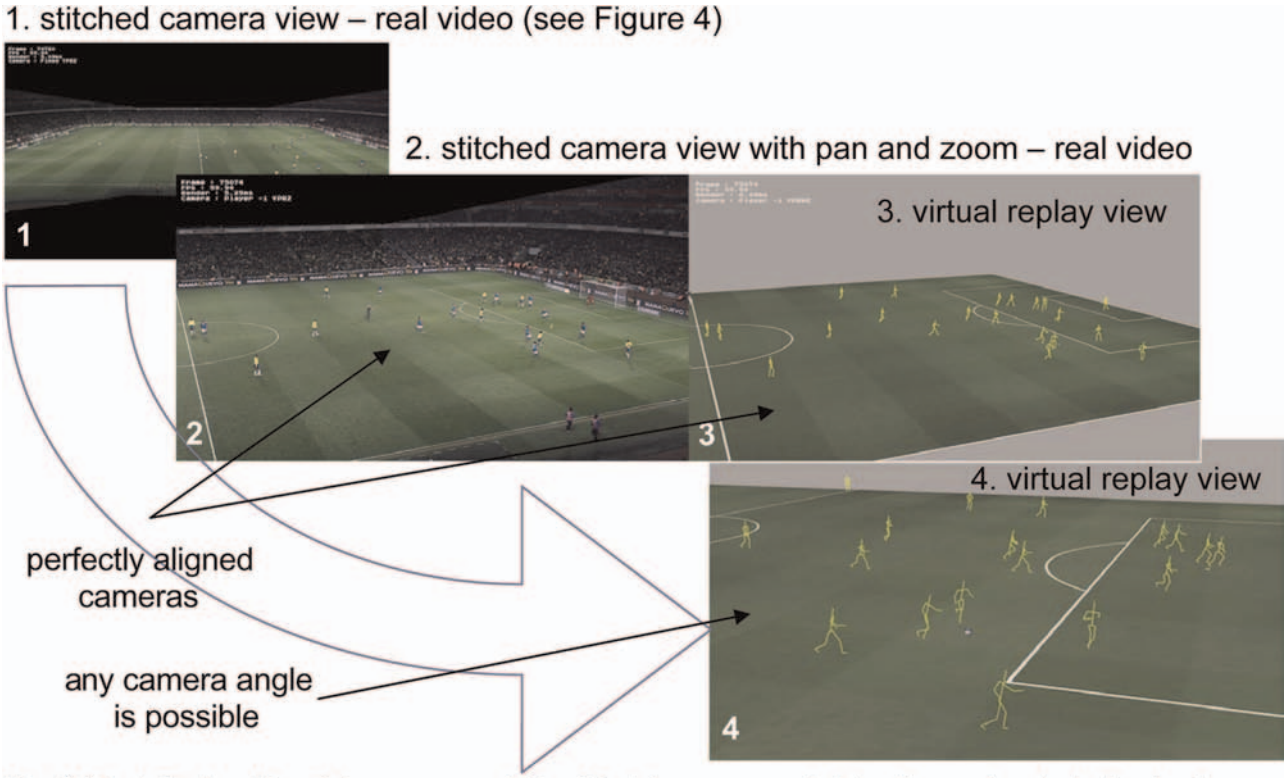

Figure 1 *Real to virtual world sequence*

The project partnership between Sky Sports and Sony provides the complimentary expertise required to deliver the vision of this project: Sky Sports' operational expertise and Sony's acquisition and image processing technology provide all the necessary components.

This paper summarises the overall system concept and architecture. The major components of the system are explained followed by a brief summary of the intended broadcast operation.

2 System concept and architecture

By means of tracking the position of all football players, referee, linesmen and ball it is intended that a sophisticated computer football rendering engine will be able to faithfully reproduce any desired sequence of play in the virtual world. This is referred to as virtual replay in this paper.

The concept of the system is to merge the real video of the football match seamlessly with the virtual replay to convince the viewer that any camera angle and view of the match is achievable.

The overall system architecture consists of a synchronous acquisition component and an asynchronous media network system (see Fig. 2).

2.1 Acquisition

The acquisition system has been designed to fit alongside a traditional broadcast camera set-up. This allows these cameras to be installed and set up at the same time as the broadcast cameras. The camera outputs are fed along with the broadcast cameras into the outside broadcast facility and are remotely controlled in the same way as all other broadcast cameras. The acquisition system cameras are fixed in a static arrangement and therefore do not require a camera person to operate. The output of these cameras are also presented to the vision mixer where they can be used as additional camera angles for direct broadcast and video replays.

The acquisition system consists of five HD cameras positioned along the broadcast gantry within the stadium (see Fig. 3). This configuration has been chosen to enable the system to be installed in the majority of major football stadiums.

The central camera cluster consists of three cameras arranged to allow the video outputs to be stitched together to produce a seamless high-resolution view of the entire football pitch (see Fig. 4). This view is used for on-air presentation and by the commentary team as it provides a complete uninterrupted view of the entire pitch. The resolution is such that all players can easily be visually

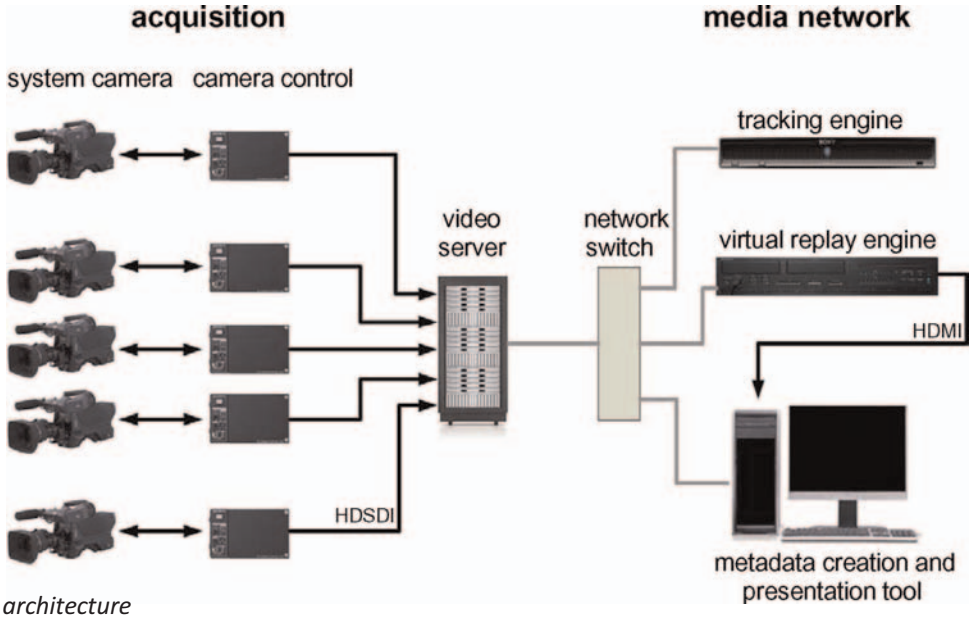

Figure 2 *System architecture*

identified no matter where they are on the pitch during play. This output is generated in real-time with a delay of approximately 1 s. For more information on how the stitched camera view is generated please see [1].

Each of the two 18-yard position cameras are arranged to capture at least half the area of the pitch. Various pairs of overlapping cameras can be formed using this configuration to allow triangulation of the objects required for tracking.

Each camera used here is a HD broadcast quality system camera. The five cameras are monitored and controlled by a camera engineer throughout the match to ensure that they remain closely colour matched and so forth. All cameras are locked to the broadcast reference and the output of each of the cameras is captured to a video server (see Fig. 2).

The video server continuously records the camera outputs for the duration of the match including a period of time before kick-off to allow the cameras to be calibrated. The five 25 psf video streams are stored as MPEG2@50Mbps in 422P@HL. All other components of the system connect to the stored video data using gigabit Ethernet. Access is non-linear and can follow the live capture with a minimum latency of approximately 1 s. This video server was designed and built specifically for this project.

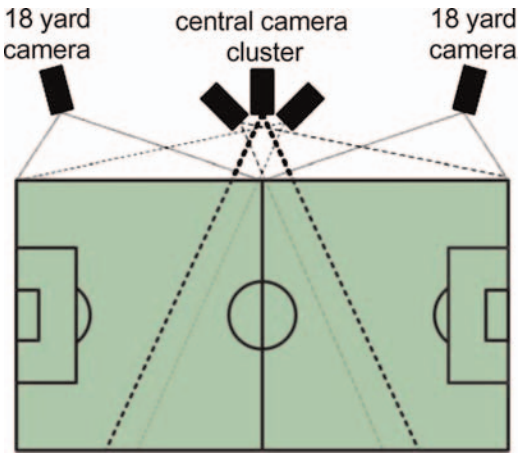

Figure 3 *In stadium camera positions*

2.2 Media network

The video server provides asynchronous and non-linear access via gigabit Ethernet to the stored video and any subsequent football metadata that is generated. This enables the analysis of the football match to be completed during live operation or post-match. There is no distinguishable difference between live and post match processing, which simplifies the system to one mode of operation.

The media network is designed for flexibility and scalability to allow multiple clients to access the video and data. However, for a match day presentation only three additional components are required: metadata creation and presentation tool; tracking engine and the virtual replay engine (see Fig. 2). These three components are presented in the following sections of this paper.

3 Metadata creation and presentation tool

The metadata creation and presentation application is a PC-based tool and is the only component in the system that requires an operator. The tool is capable of decoding and displaying all five camera views in real-time with non-linear access and trick play to any instance in the match. All generated metadata is overlaid on the video providing the operator with a rich graphical user interface (GUI) (Fig. 5).

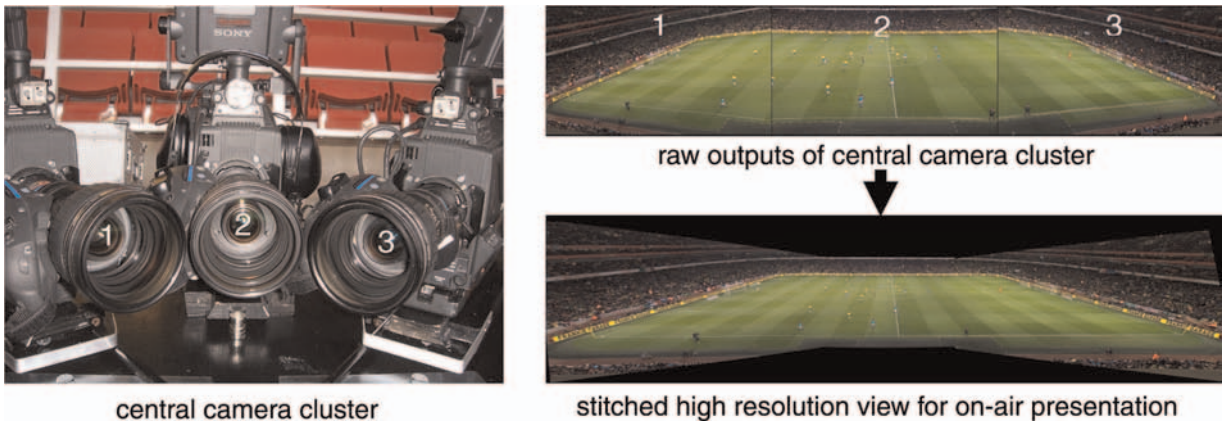

Figure 4 *Central camera cluster and stitched camera view*

3.1 Pre-match data

Pre-match data such as team names, player names and player positions are held on a database. This database holds current information on all aspects of many leagues around the world. This makes selecting the initial team line-ups, including substitutes, straight forward as no individual player names need to be typed in. Team strips, official strips, the stadium, time of day and weather conditions are also selected as this information is required by the gaming engine employed to enable the correct assets to be loaded for rendering.

3.2 Camera calibration

Camera calibration is required to determine the real geometry of the scene from each camera view. The calibration data allows the system to determine the real position of tracked objects such as the football players and the ball relative to the real dimensions of the pitch. In addition, the calibration data are also used to stitch the three images of the central camera cluster to produce a seamless high-resolution view of the entire pitch.

Calibration of each camera is achieved via a GUI, which allows various points on the pitch to be marked. Each point marked requires a correspondence to a mathematical model of the pitch. The only information required for the model is the actual dimensions of the pitch.

At this stage an additional process of marking up the boundary lines of the pitch is required to allow lens distortion of each camera to be calculated. Lens distortion data are required to allow the correct translation from the two-dimensional camera view of the pitch to the three-dimensional (3D) model of the pitch. The lens distortion calculated is also removed from the three central camera views when producing the high-resolution-stitched output. This is important as we need to match the camera of the stitched view with the camera of the computer-generated

Figure 5 *Metadata creation tool GUI*

view. The computer-generated model does not allow for lens distortion.

Each camera is calibrated and corrected independently from each other using the same pitch model. The results of the calibration are checked by projecting all camera views onto the model of the pitch. The football line markings are used to check for registration and alignment accuracy.

3.3 Football metadata creation

The metadata creation and presentation tool is capable of creating 3D player, referee, linesman and ball positional data on a per frame basis by means of an operator manually entering the data via a GUI. However, this is a time consuming exercise. To speed things up the tool has access to a tracking engine that will automatically track everyone on the pitch including the ball. From the operator's point of view the tracking happens seamlessly with the results presented in a graphical form overlaid on the real camera views. Each player is highlighted with a box and the ball is given a trail.

The operator is responsible for identifying each player by name; however, this is only required at the beginning of the sequence as the tracking engine is able to cope with player occlusions. The system does however guess the detected player names based on their location on the pitch. For longer sequences player verification is necessary although the system will alert the operator when two or more players may have had their names swapped.

Additional event data are required by the employed rendering engine to faithfully reproduce the selected sequence. Player events such as kick, trap, throw-in and header are currently manually entered via the metadata creation tools GUI.

Once a sequence has been marked up it is simply exported to the virtual replay engine for verification and on-air presentation.

3.4 On-air presentation

The on-air presentation output of the system is based around the stitched high-resolution view as generated from the central camera cluster (Fig. 4). From this high-resolution view a new HD camera view can be generated allowing the operator to perform zoom pan and tilt. This camera movement can be achieved independently of the playback of the video.

It is this generated camera view that forms the basis of the on-air presentation. Telestration graphics are rendered on this camera view. All graphics are rendered using the real geometry of the scene giving the illusion that the graphics are painted on the pitch. It is this camera view that is simultaneously and precisely matched in the virtual replay engine. As the operator pans and zooms the new HD camera view the virtual replay camera also pans and zooms at the same time. The operator can then mix between the stitched view and the virtual replay view at will. Once in the virtual replay view this camera has the added capability of being able to fly to any position. Also the camera parameters can be altered to obtain a different feel of shot. Once the desired progression of play has been achieved the operator (at the press of a button) can trigger the return of the virtual replay camera back to a position that tallies with the stitched camera and a mix is performed back to the real video. This gives the viewer a context that leads from real to virtual and back to real (Fig. 1).

4 Tracking engine

The tracking engine resides on a cell broadband engine-based server platform that operates as a co-processor and is present on the media network (see Fig. 2).

The tracking engine is simply passed the start and end frame number and location of the video sequence required for processing. The tracking engine retrieves the video data from the video server and automatically processes the selected sequence. The position data of the players, referee, linesmen and ball with the associated frame number are streamed back to the metadata creation tool on a per frame basis.

The engine automatically detects the initial player positions and categorises them into teams, goal keepers and officials. Ball tracking is achieved by simultaneously tracking several objects that have a high probability of being the ball. All these ball objects are presented to the operator for correct selection. The tracking is robust to adverse and varying light conditions and can be configured to work with any number of cameras.

The tracking engine is scalable to allow multiple engines to coexist on the same media network. For higher performance systems multiple engines can be configured to work together.

5 Virtual replay engine

The virtual replay engine is responsible for creating a high quality rendering of the chosen sequence of play.

The virtual replay engine operates as a plug-in to the metadata creation and presentation tool (see Fig. 2). The virtual replay engine is controlled via an extensive application programmable interface, which enables complete control over camera parameters, camera position and also the playback position of the sequence. This enables the presentation system to match the real camera view precisely with the virtual replay camera view to create a seamless mix between the two. The mix can be initiated for frozen frames as well as full motion sequences.

It is recognised that for absolute accuracy, full body motion capture of every player on the pitch would be necessary. However, this hybrid approach has yielded excellent results and remains one of the most exciting areas of the continuing research aspect of this project.

6 Broadcast operation

The system is used to provide the commentary team with a complete view of the football pitch for the duration of the match. The commentator will also be able to scrub the video to review recent incidents as and when they occur.

The system's main output are the generated virtual replays for half time and full time analysis. The system can also generate the linesmen's and referee's point of view for specific incidents during the course of play.

Various short sequences of play are indentified during play for the creation of a virtual replay. The operator produces the various replays via the metadata creation tool. The replays are then loaded into the presentation tool for on-air presentation.

7 Conclusions and future work

We have presented a summary of the overall system for producing virtual replays of a football match as an enhancement to the broadcaster's sports graphics system.

The key features of this system are the ability to seamlessly mix from a real camera view of the match to a computer-generated virtual view of the match; the ability to plug in a football rendering engine such as a high quality games engine; the creation of a stitched high-resolution view of the entire pitch and the ability to easily store all the video and data for the entire match for instant recall when required.

We have also shown that the system has been designed for ease of installation and integration into a live broadcast infrastructure although larger more permanent systems can also be catered for.

This system has been trialled at various football grounds such as the Madejski Stadium, Old Trafford and the Emirates Stadium. The latest successful trial as of writing this paper was the Brazil against Italy match held on 9 April 2009 at the Emirates Stadium.

Future requirements of this system include features for football analysis such as the simulation of alternative match scenarios. It is this type of interactivity that could be presented to the viewer as a downloadable data package to his or her games console to produce endless alternative scenarios or to take over the control of a certain player at key instances of the match to see if they can influence the result. This is just one of many potential uses of this metadata.

Several new research areas are being considered to improve realism and speed of mark-up. For example, the automatic extraction of pose and gait for each player can be used to align the virtual world more closely with the real world. The automatic extraction of events such as left kick, header and so on will reduce the manual mark-up effort and thus reduce the time to create virtual replays. Another research area would be to produce 3D virtual replays suitable for 3D broadcasting.

Finally, we have demonstrated that we have made a giant step of getting closer to the vision of creating a photo realistic, high realism virtual football match in near real-time to enable the coexistence and fusion of the real world with the virtual world.

8 Acknowledgments

The authors thank colleagues in Sony BPRL, Sony PSE and Sky Sports for their support and contribution to this work. We also appreciate the encouragement and continued commitment to our research from our colleagues in Sony Corporation, Japan.

9 Reference

[1] PORTER R., BERESFORD R., WAGG D. S HAYNES: 'Sports content with intelligent image processing'. *Proc. 2007 Int. Broadcasting Convention,* 2007

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

The truth about stereoscopic television

D. Wood

European Broadcasting Union, Geneva 1218, Switzerland
E-mail: wood@ebu.ch

Abstract: Stereoscopic television can be regarded as a limited subset of 'natural vision'. It can provide an exciting viewing experience, but has constraints. Reviewed here is the relationship between the process of natural vision, the 'object wave' and stereoscopic television. The main elements of the creation of space image points from screen image points in stereoscopic television are outlined. From this, examples of the limitations of stereoscopic television are described: the potential convergence–accommodation conflict, infinity separation, viewing distance and depth resolution. Conclusions for programme production and the future use of stereoscopic television are drawn.

1 Introduction

Stereoscopic television is the hottest subject in media technology today. The 'word' is that it will take the home viewing experience to a whole new level. Many of those who see it for the first time, in a well adjusted and professional form, are thrilled by it. Engineers would just love a good three-dimensional (3D) TV system – after all, our business is to lead the way on the long march to greater realism.

But this is also the time for truth and honesty – not wishful thinking. The truth about 'stereoscopic television', as we know it today, is that it does not provide the complete natural experience of viewing the world. It is actually a 'sub-set' of it, which has many constraints and limitations, and a tendency to eyestrain. It will certainly be used, but we need knowledge and creative skill to minimise its shortcomings.

Stereoscopic television can be seen as a 'first generation' 3D TV (in the ITU-R WP6C, 3D TV is seen as likely to evolve in three stages, beginning with stereoscopic television and ending with object wave recording, explained later in this paper and in appendix) system. The parameters of stereoscopic television have to be chosen based on a set of 'average' circumstances, not absolute circumstances, so there will always be those for whom the system does not work well. The production grammar that is compelling here is more limited than for normal 'planoscopic' television, and sometimes different. There will be new vocabulary to absorb, new knowledge to gain and training to be done, if we are to move beyond throwing battle axes at the audience, and giving them eyestrain after an hour.

Engineers at least need to understand how 'stereoscopic' television fits with the 'natural vision' process, where to look for its limitations, and to continue the search for even better 3D TV systems for the future. The purpose of this paper is to introduce these issues.

First, we need to consider what the 'natural vision' process is, and how stereoscopic television is derived from it.

2 Process of natural vision

What is 'natural vision'? What are the processes that occur when we look at the world around us?

There are two inter-related stories to tell – briefly in this short paper. First is what happens to light on its way to our eyes – the story of the light's path. The second is what happens in the eye and brain when it captures its images – what is termed 'cyclopean' image creation. When you understand what they do, both seem nearly incredible, and surpass in complexity much, or all, human engineering achieved to-date.

2.1 Light path

In the natural world, light coming from nature (the sun), or an artificial light, strikes an object and is partly or totally

absorbed or reflected. The colour of an object we see is the particular wavelength in the illumination that the object cannot absorb. Light can also be transmitted through objects that have a degree of transparency. The reflected or transmitted light normally sets off 'in all directions', and for every object it is the same situation. What is passing towards us, when we see something naturally, is the sum of all the refracted or transmitted light waves that are generated by the combination of light source and objects (Fig. 1).

The totality of the waves is also a wave, and it is termed the object wave (*O*). Imagine that you have an empty picture frame, and you hold it up in front of your face. What is passing through the frame, and strikes (very lightly!) your face is '*O*'. Like all waves it too has amplitude, frequency and phase. '*O*' is the integration of all the myriad scattered waves generated by the objects before you.

A 'planoscopic' camera takes one sample from the object wave at a particular point in space. For most people, our two eyes each take 'samples' from the object wave, about 6.25 cm apart. Our eyes and brain use the amplitude, frequency and phase of the samples from the object wave, and the way they change when we move, to interpret the scene that lies ahead of us. This latter process is outlined in section 2.2.

If we could devise a recording system to capture and reproduce the 'object wave' for a plane in a given area in space, we could reproduce 'natural vision', with all the features of natural vision – focussing, converging – and it would be eyestrain free. This should be our ultimate goal for television. Systems such as 'multiview' and 'IntegralTV' (these may be considered as second generation 3D TV systems) lie between the stereoscopic and the complete object wave recording. They are (arguably) approximations to an object wave.

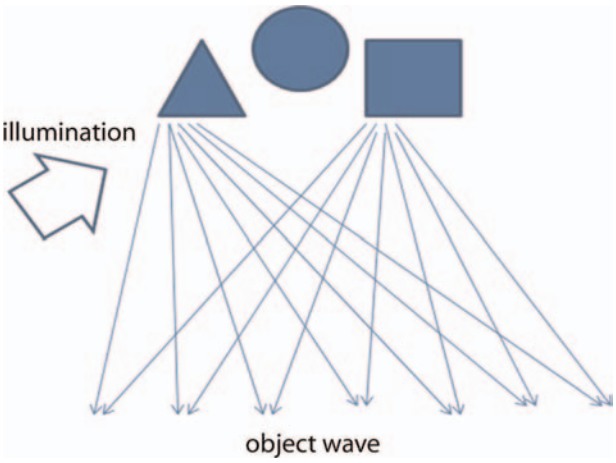

Figure 1 *Light is reflected from all points of all un-obscured objects, and the totally of waves forms an object wave (O)*

The 'hologram' is a first (but surely not the last) method of recording a limited type of object wave using a photosensitive film. Remember that the job is to capture the amplitude, frequency and phase of a wave over a given area, on a reproducible recording media. A hologram does this by simplifying its task, by working at one wavelength only, and by combining phase and amplitude into one variable. This makes for a recordable signal. More details are given in the appendix to this paper.

2.2 Processes in the eye and brain

Each eye focuses an image 'sampled' from the object wave onto the retina at the rear of the eye. The image available on the retina is inverted, and somewhat distorted because of the curved rear surface of the eye. The brain has to put all this right – but it has very good software to do so.

The brain then lays one image on top of the other, and derives from that a 'fused' image that combines the two, but has perceived depth, and which is projected forward in the mind's eye view of the scene. This image appears to be being sensed from the centre of the head, and hence is termed the 'Cyclopean' image. The (marvellous) process is termed 'stereopsis'.

There are two partners in the process: accommodation (or focussing action by the muscles in the eye) and convergence (or muscles pointing the eyes at an object). These are separate actions, but in natural vision circumstances they are done in concert. They are done on instruction from the brain, which bases its request on what it sees in the picture – the depth cues.

These cues can be derived from 'planoscopic' elements in the picture such as the relative size of known objects, perspectives, occlusion (shielding of one object by another), textures, contrast and shadow, and from the 'picture-differences' element – the differences between the left and right eye pictures. This 'disparity metre' functions something like a camera rangefinder – for those old enough to remember what that is. The amount of eye-muscle effort is a clue for the brain about how far away an object is.

There are many other elements of the process that influence the final '3D' picture we see naturally, such as integration time and eye receptor sensitivity, and there is also more research to be done, but these above are the main elements.

3 Stereoscopic television process

The main features of a stereoscopic television system are that two cameras are used that are set up and aligned with distance apart appropriate to the scene being shot, and the producer's intention for it – but also bearing in mind that the final

pictures will be viewed with human eyes at about 6.25 cm apart.

The cameras loosely parallel a viewer's two eyes, although there are many circumstances when they do not have the same spacing or 'convergence' as human eyes. The two images (L, R) are edited and delivered by a broadcast channel or other means, and are taken to be the images for two human eyes viewing at their normal spacing. These are two 'samples' of the object wave with spatial parallax. They can be 'fused' in the mind (a proportion of the public, possibly between 5 and 8%, is not able to fuse such images, because the processes of accommodation/convergence do not work as they do for natural vision), and the result is a single picture with depth. However, no amount of head movement can produce any change in the views, as it would in a natural vision situation, and as a recording of an object wave would do.

In the viewer's display, the two images are arranged to be accessible only to the corresponding left or right eye of the viewer. This is done by 'active' or 'passive' means.

In the active systems, there is time multiplexing of the left and right pictures in the display. In the passive systems there is colour range separation of the L and R in the displayed picture ('colour anaglyph'). The viewer has to wear special glasses which restrict access of the intended picture to the correct eye. Current BluRay and DVD use colour anaglyph; and, because it does not need a new TV display, it is the usual method for stereoscopic broadcasts made to date.

The active systems can use 'shutter glasses', or a combination of switching in the display the polarisation plane of a screen front filter in alternate frames, coupled with polarised glasses with corresponding different polarisation planes in each eye.

Although each approach has advantages and disadvantages, future broadcast systems will probably be based on an active system, which gives superior quality to anaglyph, because of improved colour rendering, lack of ghosting and often better resolution.

There are a range of proposals for constructing the delivered combined left and right eye signal for active systems. These include providing both signals at full HDTV resolution, half horizontal, or vertical, or diagonal resolution, or providing one HDTV signal plus a difference signal, or an HDTV signal plus a 'depth map' (this is useful for a multiview service with more than one pair of images).

It is important to note that these two signals recorded and provided are planar images from two cameras, and there is no 'phase' information for the eyes to use. Accommodating the eye, based on the brain's assessment of distance from the disparity cue, does not produce a sharper image on the retina. There is only one place to obtain the sharpest image, and that is the plane of the screen. This is explained later in this paper.

We begin now to see the range of limitations on stereoscopic systems compared to 'natural vision'.

We can go further by considering the photograph of an L and R picture given below. In the photograph, the two side-by-side stereoscopic pictures, L and R, are similar but not identical. If you look carefully you will see that specific objects or edges are more to the left in the L picture than in the R picture. This is because the viewing position of each eye is slightly different. We call the same objects in this stereoscopic world 'homologous points'. The physical space difference between homologous points and the edge of the frame varies with the depth of the object in the scene.

The picture, taken at the motor museum in Brussels not the author's garage, shows greater disparity of homologous points, with greater depth in the picture

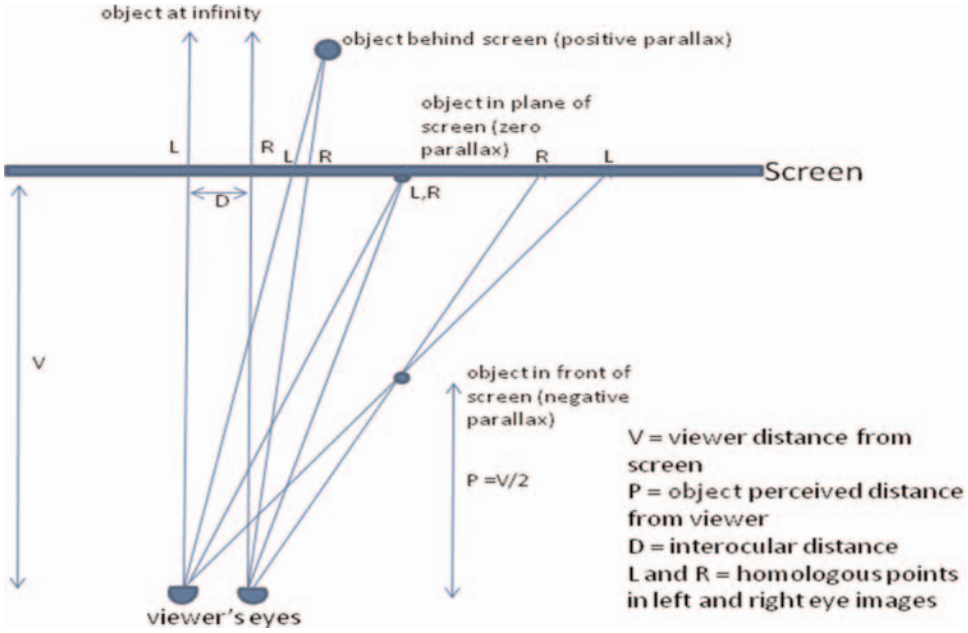

Figure 2 *Schematic of the construction of space image points from optical image points on the screen*

The viewer's eyes have a lateral separation D, are equipped with selecting spectacles, and view homologous points L and R for different objects in the scene. The difference in the position of the L and R creates perceived distances as shown, because the disparity seen is the same as would be seen if the objects were really at the positions indicated. This is the basis of the illusion that is stereoscopic television.

Fig. 2 above is a representation of homologous points of different depths in the scene. On the right are two homologous points L and R associated with an object intended to be perceived to be located half way between the viewer and the screen. Moving to the left, next to that are the two homologous points L and R associated with an object in the plane of the screen, which are 'co-located'. Next to that are two homologous points of an object behind the screen. Finally, we take the case of an object at 'infinity', where the lines of sight for the two eyes are virtually parallel. Notice that the order of the left and right homologous points switch over when we pass through the screen plane.

Understanding the 'mechanics' of apparent depth from this figure, we can deduce many of the limitations on stereoscopic television.

Our eyes and brain are an automated system that will always try to give the brain the most coherent picture it can of the scene in front of it. Eyestrain arises when our eyes are asked to do things they do not do normally, which can include an abnormal consequences of accommodation and convergence, difficulty in fusing or focussing, changes in the depth settings of the scene, which are more rapid than normal, and having abnormally different pictures in the two eyes.

Our eyes and brain can be tolerant of abnormality, but after a time it will 'rebel' and ask you kindly to take the glasses off. By good practice, careful choice of parameters, formats and viewing environments, we can 'minimise' the tendency to eyestrain, and maximise the eyestrain-free viewing time.

Here are some (but by no means all) of the limiting elements that we need to be aware of, and which call for judgements about 'average' conditions.

4 Accommodation–convergence conflict

As can be deduced from Fig. 2, while the eyes may 'point' towards the apparent position of the object ('convergence'), except for the case where the object is actually in the plane of the screen, focussing ('accommodation') on the apparent position of the object will not produce the sharpest image on the retina. The eye actually needs to focus on the plane of the screen, which is where the 'real' optical image lies.

Thus there can be, to varying degrees, eyestrain because the eye–brain can be asked to do something it does not do normally, and 'separate' accommodation and convergence. The potential for eyestrain can be reduced if the main action in the scene takes place in the plane of the screen, or if the viewer is sufficiently distant from the screen for both apparent object and screen to be within the depth of sharp focus of his eyes.

Because viewers are more physically distant from cinema screens than from television screens, this conflict may arise more in the television environment than in the cinema environment.

5 Infinity separation

As seen in Fig. 2, objects in a scene which are distant should have an on-screen separation of the same distance as the viewer's interocular distance D. If this is the case, the viewer's gaze will be straight ahead, as it would normally be when looking at distant objects. If this separation is actually greater than the interocular distance, and the eyes have to point outwards for distant objects, the L and R can be difficult to fuse – and there can be eyestrain. Making the infinity separation shorter will bring infinity forward in the picture, and contribute to the 'playing card effect'. Heads you lose, tails you do not win.

However, the choice of what exactly to make this infinity separation distance in a TV broadcast is not obvious for two reasons.

First of all, there is a variation among humankind in inter-eye spacing. The average distance is 6.25 cm, but people actually vary from 50 to 70 mm. We can only choose an average, and hope the disturbance is not too great for others.

Secondly, the broadcaster has (in theory) to know exactly what screen size the viewer is going to use to adjust the infinity separation to 62.5 mm (if this is the safest bet). This is an absolute number, and needs to be therefore independent of screen size. But how can he know, before he broadcasts, exactly what screen size the viewer will be using? You see the problem? In future could we develop displays that can be made adaptive so that the infinity separation is always corrected (by moving the two pictures L and R together or apart in the set itself) for 62.5 mm? Until then, we are only in the business of providing for a Mr and Mrs Average and hoping.

6 Ideal viewing distance and geometrical congruence

For planoscopic television systems (LDTV, SDTV, EDTV, HDTV) there is a 'design' viewing distance that was used to identify the amount of resolution needed in the picture in each case. The 'design viewing distances' are 8H, 6H, 4H and 3H (see ITU-R BT 1127). TV systems are designed to saturate the eye with detail at the design viewing distances, based on the resolution limit of the eye (this is usually taken to be 1/1 min of arc). All this goes topsy-turvy for stereoscopic television, because we have the dimension of depth to consider too.

As can be seen from considering Fig. 2, the apparent front to back depth that objects have in the perceived scene is not just determined by the difference in spacing of homologues points on the sensor and on the screen; it is also determined by the absolute viewing distance.

Having a homologous point disparity of, for example, −D (as in Fig. 2) means that $P = V/2$, whatever V is. In other words, the further back you sit from the TV, the more elongated objects will become in the depth direction. In other words, the 'geometrical congruence' of the scene on the screen with the scene it represents is influenced by viewing distance. So, where should I sit?

There is even more complication because geometrical congruence is also influenced by the camera lens focal length used in shooting the scene. As portrait photographers know, long lenses flatten objects and short lenses lengthen them. Ideally, for things to look 'right', we might use a camera lens for stereoscopic television that is quite wide – like the eye's lens – although that too will vary with individuals, because our eyes vary in size. But a short lens is not what you want to use when you cannot get close to the actors or action.

The ideal viewing position for stereoscopic television, which is called the 'orthoscopic' position, is going to be difficult to find. It can be done. Imagine that you are sitting in front of the stereoscopic television set, and you had a camera with you with a lens the same as was used to make the stereoscopic programme. Now, you just move your seat until the screen fills the camera viewfinder. The two angles of view – taking and viewing – are now the same. You have the best seat in the house! The problem is that – curse the cameraman – the camera focal length will probably change. Different programme content needs different focal lengths, and they can be changed during a programme. The ideal viewing seat moves with every change of focal length.

7 Depth planes

One of the principle defects of stereoscopic systems can be 'the playing card effect' when the picture appears to be made up of flat objects. Different factors can cumulate to exacerbate the effect, and great care is needed.

Part of the reason is that in a stereoscopic television system, depth detail itself is usually limited anyway. The detail in the original z (depth)-direction is transformed to the horizontal direction. The combination of the L and R images must provide detail in all three planes; x, y and z. This depth detail is contained in the difference in the L and R pictures. By examining Fig. 2, we can see that the depth detail in the z-plane will be limited by the detail possible in the equivalent portion of the on-screen horizontal deviation.

The number of depth planes possible between the plane of the screen and the rear infinity plane behind the screen will be constrained by the resolution possible in 6.25 cm of actual screen width. The limit on definition between the screen plane and half-way to the viewer will also be the resolution possible in a screen width of 6.25 cm.

If the screen is 100 cm across and there are 1920 samples per line, the maximum number of depth planes between the

screen and infinity will be limited by $6.5/100 \times 1920$, multiplied by a fraction that is the equivalent of the Kell factor. Certainly much less than the horizontal or vertical resolution.

The limited number of depth planes is not the only element effecting 'cardboarding', but it needs to be a factor in the equation of careful production.

8 Programme production limitations

Although it will always be technically possible to make stereoscopic content which is '2D compatible', in that a two-dimensional (2D) picture is available, 'creative compatibility' may be not always possible. What makes the 'best and most compelling' programme shots in 3D may not be those for 2D. In the end, we have to 'sell' this system to management and the production community. Below are some of the production grammar limitations for stereoscopic television. How much of a work of art will a 2D version with these constraints be?

- Scenes shot in stereoscopic TV should be all in focus from front to back. Occasional shots can be made with a smaller depth of field, but they are harder for the brain to fuse.

- To produce a shot where the depth is clearly evident, the ratio of inter-camera separation to the principle object in the scene should be 1/30 to 1/50, depending on the focal length of the camera lens.

- In the 'hand over' between shots in the final edited programme, the position in the z-plane of the principle subject should be similar at the scene transition, to reduce eyestrain.

- Active stereoscopic systems use time multiplexing and thus do not present the same motion phase at the same time, and thus care will be needed in motion portrayal (with events such as sports) to prevent jerkiness.

- Cameras with small focal lengths produce greater geometrical congruency to the scene shot – although creative use of longer focal lengths may however be part of a journey of discovery of creative uses of stereoscopic television.

- Stereoscopic dramas usually have special scripting of 3D effects.

9 Conclusions

Stereoscopic television will certainly be a part of the media landscape in the years ahead. The 3D effect can be powerful and moving, and adds considerably to the viewing experience. But stereoscopic television is a limited system, and without care can lead to eyestrain. We need to make sure that eyestrain does not 'poison the water' for the public.

Stereoscopic services will surely be available – at least if the world's technical community can agree standards – but they will not replace conventional television, because it has its own advantages and creative space.

There is a job to do for industry. Alignment of stereoscopic L and R requires precision adjustment with 1–2 pixel accuracy. Test charts and automatic software systems will be needed for programme production and post-processing, to avoid an army of technicians being needed for programme production.

Stereoscopic television must not be 'forced' quickly on the production community or the public, or it may fail, as it has done in repeating cycles in the cinema. The industry and the public deserve well considered standardisation of systems and guidelines for production.

But in the end, it is not the job of the engineer to make life difficult, but rather to make things possible. Let us do that, but keep honest and our feet on the ground. As they say – 'stay real'.

10 Acknowledgments

The wealth of knowledge available on stereoscopy is due to many pioneers, including the late Charles Smith, and Nicolas Lodge. Their legacy continues today with the many members of ITU-R WP6C, which is studying 3D TV.

11 References

[1] MESSERSCHMID U., SAND R., WOOD D.: 'Relief in sight?', *EBU Rev.*, 1986, **XXXVII**, (6)

[2] DOSCH C., WOOD D.: 'Can we create the Holodeck?', *ITU J.*, 2008, (9)

[3] LIPTON L., NOSTRAND V.: 'Foundations of stereoscopic cinema' (1982)

[4] SCHREER O., *ET AL.*: '3D video communication' (Wiley, 2005)

[5] SMITH H.M.: 'Principles of holography' (Wiles-Interscience, 1969)

[6] SPOTTISWOODE R., *ET AL.*: 'Basic principles of three-dimensional film', *J. SMPTE*, 1952, **59**

12 Appendix

12.1 Elements of holography

A hologram begins by removing the variable of wavelength, by using only one as the source of illumination of the object being photographed – coherent monochromatic light from a laser ('R') (although colour primaries can be used to create a colour picture). Then it uses a way of folding-in the phase information into the amplitude, in such a way that they both can be recovered later.

Image sensors do not record both amplitude and phase as such, they record 'irradiance', which is: $|O|^2$, so exposing a sensor to 'O', without doing anything else, just produces a fog on the plate (as photographers know well). To obtain a sharp image on the plate, we need a small aperture and/or lenses, which samples O down to a single planar wave. This conventional image so produced is fine for what it is, but is not what we need for 'natural vision'. The trick needed in the hologram is to record the amplitude and phase of the object wave, over an area.

The hologram process needs thus to include a function like 'modulation' in broadcasting – to transform the wanted signal to a form that can be used. In this case, the 'modulation' is done by adding a second reference light wave R at the same wavelength as the object wave illumination, at the photosensitive film surface. The plate records the sum (or interference pattern) of O and the reference wave R.

Thus, what is recorded is thus: $|O+R|^2$, and, by expansion: $|O+R|^2 = |O|^2 + |R|^2 + OR^* + O^*R$ (* denotes conjugate complex).

Now, happily O can be recovered by illuminating the plate with R, and this can be shown as follows

$$R|O+R|^2 = R|O|^2 + R|R|^2 + |R|^2O + R^2O^*$$

The third term here is $|R|^2O$, which is a constant times 'O', and thus, voilà, we have now recovered the object wave.

This process is comparable in some ways to radio frequency modulation and demodulation. Future generations of engineers will surely develop much more sophisticated ways of transforming light into recordable signals, just as they have for baseband signals in broadcasting over the last 100 years, or of finding recording media which record both amplitude and phase. Until they do, we need to use a simpler process for any '3D TV', stereoscopic television, but we must still encourage research and development of 'Full3D TV'.

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Interview

This year the IET and IBC have identified an outstanding paper from young professional Thomas Jaeger, a PhD student at the University of Dortmund, Germany. With the help of fellow PhD student, M.Y.Al Nahlaoui, Thomas has submitted a paper that is the reviewers' choice to be published. Simon Yarwood, Community Development Manager at the IET, caught up with Thomas to find out a bit more about him and his paper.

Tell us a bit about yourself and what you do

I received my diploma degree from the Technical University of Dortmund, Germany, in 2005 after studying information technology with a main focus on communication technology and computer science. I then began my PhD studies, focusing on image processing, at the communication technology institute, under the supervision of Prof R. Kays, which I will finish at the end of this year.

My research interests are in the fields of image processing, 3D-TV and computer vision, with a focus on stereo image processing. One main aspect of my research is the impact of the human visual system on image processing algorithms.

Can you explain to our readers what your paper is about (in laymen's terms)?

Stereoscopic, or 3D, projection in cinema is experiencing a revival nowadays. This year, a large number of 3D movies will be shown or have recently been shown in cinemas and the content creators for stereoscopic movies have to deal with completely new challenges. To give one example, it is not possible to use exactly the same art work owing to the depth of the field. A limited field depth is responsible for guiding the eye of the viewer in the scene, as the viewer would always look at the focused image areas. When presenting a stereoscopic movie with a limited field depth the viewer would also only look at the image areas in focus and the newly obtained freedom of a stereoscopic presentation is reduced to a minimum.

The other possibility would be a completely focused image with a very high field depth. Then the viewer can focus his or her attention on every object in the 3D scene. But this leads to a new problem - when watching a normal (i.e. not 3D) movie in the cinema you normally focus your eyes on the screen. To generate a 3D viewing experience, a stereoscopic projection with separate images for the two eyes is used. As the projection for each eye is nearly the same as with a normal movie, you again focus your eyes on the screen. The apparent distance of the objects in the scene has no influence on the viewer's focus. This leads to an unnatural viewing experience as every object at every distance in the scene is in focus. Our eyes and our brain can handle this property but the viewing experience is clearly lesser in comparison to the experience when watching a real scene.

This paper addresses this trade-off and gives a solution that considers both aspects. When creating the content, the field depth is as wide as possible, resulting in a completely focused image. Based on the gaze direction of the viewer, the field depth is reduced either to a range selected by the producer or to the natural field depth of the human eye. This gives an extended 3D viewing experience.

The idea for this paper emerged during discussions with my advisor, Prof Kays, about the impact of blurriness on the 3D viewing experience.

3D is in the news more and more recently, what is it that excites you about the topic?

Now that digital revolution in the movies has successfully taken place, 3D is the next major step towards the best

possible viewing experience. To achieve this, much research and development has to be completed. The technology for presentation of stereoscopic content at home is a completely new field in the area of display technology and therefore has many challenges. A large number of different approaches have been developed and are still under development to fulfil the special needs of the home environment. It is very exciting to analyse the different technologies and great ideas in this research area.

From a personal point of view, what do you see as the future of 3D?

The desire for stereoscopic presentation of movies has a long tradition. From the anaglyph projections of the 1950s and the disappearance of stereoscopic projection to fully digital production and projection, the technology has developed to a stage that makes it ready for commercial use, at least in cinemas. Now that it is part of the digital world I won't expect 3D to disappear again. This of course depends on the acceptance of the content by the viewers. Content producers are now responsible for creating high class content for the audience's entertainment. The latest news from this area seems to show that the first steps have been successful. If, some day, holographic displays come onto the market, this would be a great step towards real 3D.

What do you see as the possible challenges in achieving this vision of the future?

As I mentioned, one of the most urgent challenges is the creation of good content, as 3D technology stands or falls with the audience's acceptance. The industry would only invest 3D if it were to see a return on investment. After establishing the technology for the consumer it would be possible to develop the technology to a point far beyond the current technology. Remember, the development of flat displays are normal nowadays for us, but were just science fiction only some decades ago.

Is this the first paper you have submitted to IBC and have you been to the conference before?

This is my first paper submitted to IBC as this is the first paper that matches the conference topics. My previous research and publications at home addressed the human visual system directly and therefore have no direct relationship to applications in the broadcasting and consumer electronics field. As 3D is one of the topics for this year's conference, I saw it as a good possibility to discuss my ideas and results with a great number of professionals during the poster session.

I went to the IBC exhibition last year to do some research into video coding products. The large number of exhibitors gave me a good survey of the market situation in this field.

Apart from presenting your poster, what else will you be doing at the conference? Are there any sessions you are particularly interested in?

I am very interested in some of the 'Technology Advancements' and 'Content Creation and Innovation' session topics. I am particularly interested in the 3D sessions of course. Sessions on stereoscopic production from the producer's point of view are very interesting as they could give an understanding into what is really used and required on the set. This understanding is necessary for research that responds to the needs of the industry and other operators.

Aside from the conference I will also have a look at the latest developments in the broadcasting field in the exhibition area.

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Gaze direction adaptive field depth reduction: boosting the 3D viewing experience

T. Jaeger M.Y. Al Nahlaoui

TU Dortmund University, Germany
E-mail: thomas.jaeger@tu-dortmund.de

Abstract: The creation and display of three-dimensional (3D) content raises new challenges, especially to producers and cinematographers. Established art work in movie and TV productions, e.g. a limited field depth, cannot be adopted in the 3D world without drawbacks. Either the field depth is reduced like in normal movies, which binds the observer to the objects in focus, or the field depth is very wide, which results in an unnatural viewing experience. Both solutions attenuate the 3D viewing adventure. Presented is a solution to this trade-off by gaze direction adaptive limitation of the field depth. Based on the gaze direction, the field depth is reduced to a predefined value or to the natural field depth of the eye. The observer can move their gaze freely around the scene, focus on every object and always have a natural viewing experience. Tests showed that this approach boosted the 3D viewing adventure for the observer.

Introduction

These days, the creation and display of three-dimensional (3D) content has to be approved by producers, broadcasters and companies involved in consumer electronics. In 2009 at least twenty-four 3D movies will be shown in cinemas, and at this year's CES exhibition in Las Vegas nearly every major manufacturer of consumer electronics presented a 3D display for home theatres. Increased profits are expected from this new technology as 3D gives the viewer additional value for money.

In this contribution we present an interesting idea for boosting the 3D viewing experience for the audience. In the classical 3D presentation we have two possibilities for image composition: as a first possibility the whole image is in focus. This gives the viewer the complete freedom to move their gaze around the image. As a drawback, this image composition leads to an unnatural viewing experience. When presenting 3D content on a stereoscopic display, the eye is focused on the display or screen. In comparison to a real 3D scene where only the observed object is in focus, the whole scene will be in focus. As a second possibility, the focus can be reduced to the main actors, just as it is often used in the usual 2D movies. This creates a more natural viewing experience but constrains the gaze of the viewer to the sharp image region. The viewer can look at blurred regions and will perceive some kind of 3D image but he or she would not be able to see any details. This second possibility is therefore rarely used as stated in Umble [1] and Maier [2].

To overcome this trade-off we present a gaze direction adaptive field depth reduction. Dependent on the gaze direction, the field depth of the movie is reduced to a range selected at post-production or to the natural field depth of the human eye. We will describe both possibilities in this paper.

To implement this new method a depth map is needed, which can be derived from the existing stereo image or, if the two views are created from one image and a depth map using view interpolation, the depth information can be used directly. We accomplished subjective assessments and can now show that the accuracy of a depth map derived from a stereo pair is high enough to realise the desired depth of field effects without artefacts.

In the next section we will describe the whole system and give an overview on the required components. Afterwards we will describe these components in detail. We implemented this novel approach, in enhancing the viewing experience, in a single user demonstration system presented at IBC 2009.

Gaze direction adaptive field depth reduction

A solution to the described problem is the gaze direction adaptive field depth reduction described in this contribution. The system gives the viewer the maximum freedom and viewing experience in the 3D space. Therefore, we developed a vector based eye tracking and gaze direction detection system. With this system it is possible to estimate the gaze direction of the observer and the eye position in space. Knowing these two parameters makes it possible to accurately calculate the observed position on the display. Based on this position, the field depth of the stereo images can be reduced to a predefined value or to the natural field depth of the human eye by 3D signal processing.

The human optical system

The optical system of the human eye consists of different diffracting elements that form a lens system. This lens system creates a reduced mirrored image on the retina. The light is refracted at the transition of materials with different optical densities in the eye and is therefore focused on the retina. In the optical system of a common camera the image distance is changed to focus the image. In contrast to this, in the eye, the refraction power and hence the focal length is changed. This results in a different relation between focus, focal length, aperture and blurriness in the eye than in a camera system. The relationship between these parameters has to be modelled to give a natural viewing experience to the audience.

The relationship between the focal length f, the image distance i and the object distance o is given by

$$\frac{1}{f} = \frac{1}{i} + \frac{1}{o} \tag{1}$$

where all refractions between the different optical materials in the eye are considered.

When focusing on infinity the focal length is equal to the image distance. The human optical system has a focal length of 22.78 mm, on average, when accommodated on infinity as stated in Schmidt *et al.* [3]. Therefore, the image distance is also 22.78 mm. As the shape of the eyeball cannot be changed, this value remains constant for every instance on different distances. For the following considerations we define the constant $i_{\text{eye}} = 22.78\,\text{mm}$. As the image distance cannot be adapted, the focal length must be changed to focus on near objects. The correlation between the focal length and focused object distance is

$$f_{\text{eye}}(g_{\text{focus}}) = \frac{i_{\text{eye}}\, o_{\text{focus}}}{i_{\text{eye}} + o_{\text{focus}}} \tag{2}$$

The focus dependency on the focal length is the first and most obvious correlation. Besides this, a second non obvious correlation exists. The diameter of the pupil also depends on the focus distance. When focusing on distance objects, the pupil diameter is larger than when focusing on near objects [3]. This dependency is called convergence reaction and leads to a larger field depth when observing near objects. A precise correlation between the pupil diameter and the focus distance cannot be specified as the pupil diameter predominantly depends on the object's brightness. To account for the effect, we therefore model this behaviour as a linear approximation between the minimum and maximum pupil diameter

$$d(o_{\text{focus}}) = \begin{cases} d_{\min} & o_{\text{focus}} < o_{\min} \\ d_{\min} + \dfrac{\Delta d}{o_{\max}}\, o_{\text{focus}} & o_{\min} < o_{\text{focus}} < o_{\max} \\ d_{\max} & o_{\text{focus}} > o_{\max} \end{cases} \tag{3}$$

where Δd is the difference between the minimum and maximum diameter and is under consideration of $o_{\max} \gg o_{\min}$. Based on this correlation, the blurriness of off focus regions can be calculated. Therefore we can use the relations in Fig. 1.

The focal length of the eye is adjusted so that the object at distance o_{focus} is focused on the retina. For every point o_2 that has a smaller distance, the size of the circle of confusion can be derived using the theorem on intersecting lines to

$$u = d\,\frac{i_2 - i_{\text{eye}}}{i_2} \tag{4}$$

Under consideration of (1), (2) and (3), this results in

$$u(o_{\text{focus}}, o_2) = d(o_{\text{focus}})\, i_{\text{eye}} \left(\frac{1}{o_2} - \frac{1}{o_{\text{focus}}}\right) \tag{5}$$

with u being the diameter of the circle of confusion. The same correlation, but with a switched algebraic sign, can be derived for objects with a greater distance. A diameter lower than zero corresponds to a mirrored circle of confusion. The plot of the circle of confusion's diameter is shown in Fig. 2. We can see that the field depth is much larger for near objects than for objects far away. Based on this eye model the field depth has to be reduced. To realise this, two components are needed. First, we must estimate the gaze direction to derive the focused image point. Based

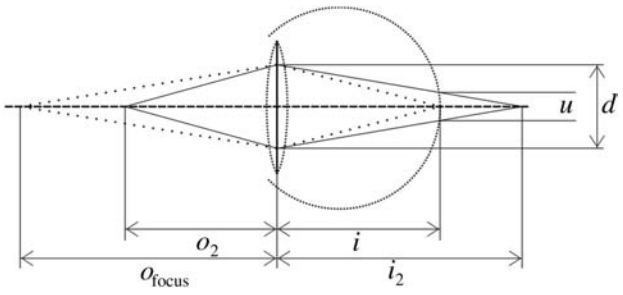

Figure 1 *Calculating the circle of confusion in the eye*

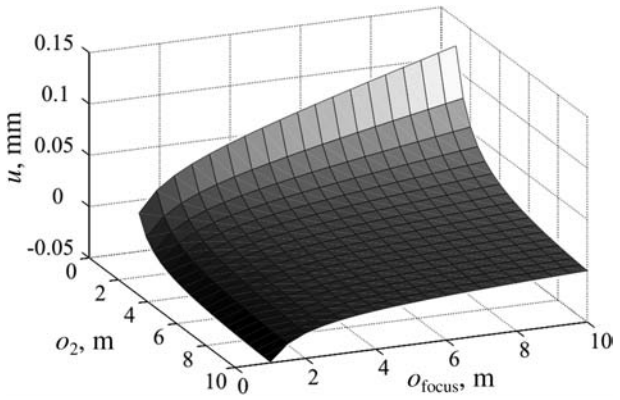

Figure 2 *Circle of confusion in the eye*

on the depth of this point we must reduce the field depth. We will describe these two components in the following sections.

Gaze direction estimation

In this approach we use a dual camera eye-tracking system (previously presented in Nahlaoui [4]) that tracks the viewing direction of a person in front of a computer monitor under normal conditions. The line of sight calculation is based on a 3D eye model. The eye position as well as its orientation is stereoscopically calculated from the images from the cameras. On the basis of known arrangements of cameras and monitors, the view point on the screen can be calculated.

System structure

The main system components are two cameras with IR pass filters and three LED clusters. The first camera with a wide angle objective (WCam) is used to analyse the whole scene in front of the computer. The position of the camera is fixed relative to the monitor. The second camera (TCam) has a telephoto lens and focuses on a limited region with one eye in the centre. Both camera images are used to extract the positions of image components needed by the view point calculation.

Owing to the sensitivity of normal CCD cameras to the near infrared spectrum (approx. 880 nm) it can be used in combination with infrared pass filters for the tracking system. This has several advantages. One of them is the reduction of possible disturbances owing to ambient light. A second one is to minimise the influence on the user and enable him to work normally. Fig. 3 shows the system structure.

In front of the monitor the WCam as well as the TCam are mounted. The relative positions should be adjusted precisely, otherwise it would produce failures in the view point calculation. In the right part of the figure some drafts of the TCam view are shown. The lower reflection (main reflection) is the reflection on the corneal surface of the

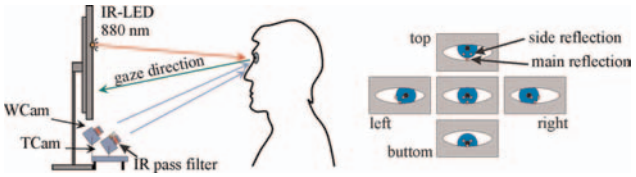

Figure 3 *System structure*

LED cluster, which is mounted on the TCam. It will be used in the next steps to calculate the eye position as well as the viewing direction. The side reflections result from the reflection of the LED clusters on both monitor sides. They are used to illuminate the scene in front of the monitor.

Functional principle

Eye model: First of all it is necessary to recall the anatomy of the eye. In Fig. 4 a cross section through the eye is shown. The main shape of the eye is given and surrounded by the sclera. At the front side of the sclera is the cornea, which is a transparent, dome-shaped window. The coloured part of the eye is called the iris. The round opening in the centre of the iris is called the pupil. The incoming light beam passes through the cornea and is controlled by the iris aperture before it can be focused by the lens on the *fovea centralis* on the retina. The visual axis differs from the optical axis. This is caused by the position of the fovea. These anatomic characteristics are used to implement the eye model in Fig. 5.

The eye is modelled by two globes. The first one, the eye ball, has a radius of 11–12 mm. The second one is the globe of the corneal curvature. The section plane between both globes is defined as the iris plane. The pupil is a small disc in the centre of the iris as well as in the real eye. The eye ball centre, the centre of the corneal curvature (CoC) and the centres of the iris and the pupil lie on the same axis. This axis is equivalent to the optical axis of the eye and defines the uncalibrated viewing direction in this model.

View point calculation: The view point calculation is done in several steps. First of all, the eye is located in both camera images to calculate the distance between the eye and the projection centre of the TCam (TPC). The direction vectors $\vec{v}_{\mathrm{T}}$ and $\vec{v}_{\mathrm{W}}$ define the direction from the camera projection centres $\vec{p}_{\mathrm{T}}$ and $\vec{p}_{\mathrm{W}}$. Mostly, it is necessary to replace the simple intersection calculation of two rays with a minimum distance calculation. Therefore,

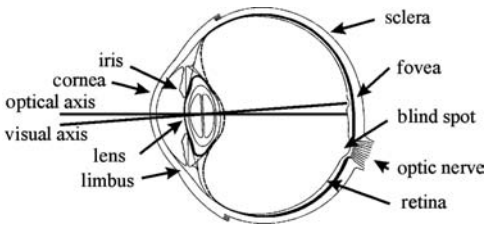

Figure 4 *Eye anatomy*

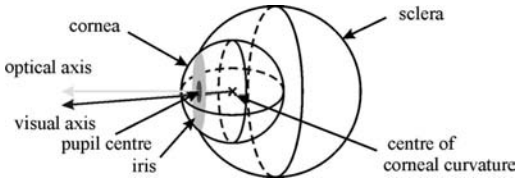

Figure 5 *the eye model*

the cross product of both directions is calculated by the following equation

$$\vec{n} = \vec{v}_{\mathrm{W}} \times \vec{v}_{\mathrm{T}} \tag{6}$$

The ray $\vec{r}_{\mathrm{T}}$ (Fig. 6) intersects a plane formed by $\vec{n}$ and $\vec{v}_{\mathrm{W}}$ in exactly one point $\vec{l}_1$. The ray $\vec{r}_{\mathrm{W}}$ can be used in the same way to get the intersection $\vec{l}_2$ with the plane formed by $\vec{n}$ and $\vec{v}_{\mathrm{T}}$. The eye position is estimated as the middle of the line connecting both intersections

$$\vec{s}_{\mathrm{eye}} = \vec{l}_1 + \frac{1}{2}(\vec{l}_2 - \vec{l}_1) \tag{7}$$

which is the point of minimal distance between both rays. In the second step, the main reflection (MR) in the TCam image is searched for. The camera parameters can be used to calculate the direction of the vector $\vec{v}_{\mathrm{MR}}$ from the TPC to the MR on the curvature of the corneal surface

$$\vec{v}_{\mathrm{MR,W}} = \begin{pmatrix} x_{\mathrm{W}} \\ y_{\mathrm{W}} \\ z_{\mathrm{W}} \end{pmatrix} = \vec{C}_{\mathrm{in,W}}^{-1} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} \tag{8}$$

where $\vec{C}_{\mathrm{in,W}}^{-1}$ is the inverse intrinsic camera matrix of the WCam, and x_i and y_i are the pixel positions of the current object in the camera image. The same equation is used with other image objects. The camera matrix should be replaced with $\vec{C}_{\mathrm{in,T}}$ to use (4) with the TCam as well.

As shown in Fig. 5, the position of the CoC can be calculated by using the TPC → MR vector ($\vec{v}_{\mathrm{MR,T}}$) with the previously calculated eye distance $|\vec{s}_{\mathrm{eye}}|$ and the radius of the corneal curvature.

In the next step, the position of the pupil centre has to be calculated. Therefore, the pupil is localised in the TCam image as well. In combination with known camera parameters, the direction of the vector $\vec{v}_{\mathrm{p,T}}$ between the TPC and the pupil centre is calculated. To get the exact position of the pupil, the intersection of $\vec{v}_{\mathrm{p,T}}$ with the corneal curvature with radius r_c is calculated. To get the real position in space it is necessary to take the refraction of the cornea into account. The final pupil position is given by an intersection with a globe with the radius of the CoC → PC distance (r_p). This procedure is visualised in Fig. 7.

With both coordinates (pupil centre and centre of corneal curvature) calculated, the viewing direction can be calculated. It is assumed to be the CoC → PC direction.

In a final step, the view point on the screen should be specified. This will be done by calculating the intersection between the CoC → PC vector and the monitor plane. To correct the offset between the optical and visual axis (as shown in Fig. 4) a calibration should be done. The horizontal and vertical offset angles will be taken into account in the view point calculation.

Field depth limitation

For the limitation of the field depth, many algorithms are known, e.g. Potmesil *et al.* [5]. For the implementation in the demonstrator we used our algorithm previously presented in Jaeger [6]. Fig. 8 shows the block diagram of the algorithm. The stereo images and a corresponding depth map are required as input to the algorithm. If a depth map is available, e.g. from a video-plus-depth video stream, it can be used directly. If a depth map is unavailable it can be derived using known algorithms, e.g. Atzpadin *et al.* [7] and Kauff *et al.* [8]. For the further processing we assume a dense depth map.

We can determine the depth information of the observed image point using the information from gaze direction detection and a depth map. This depth information is used as the focus distance. We can then calculate the blurriness

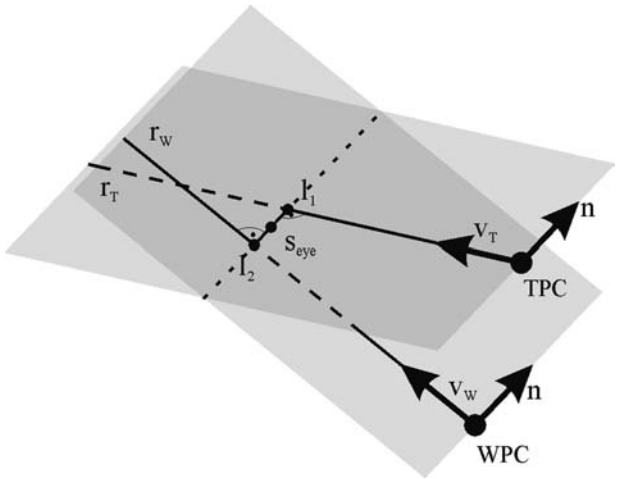

Figure 6 *Point of minimum distance between two rays*

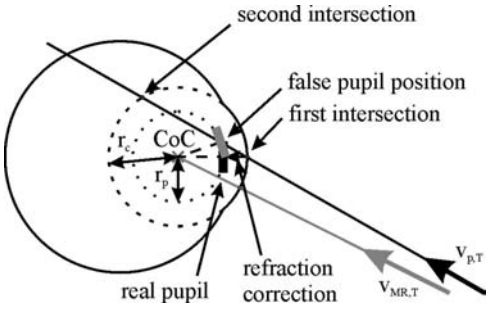

Figure 7 *Calculating the centre of the pupil*

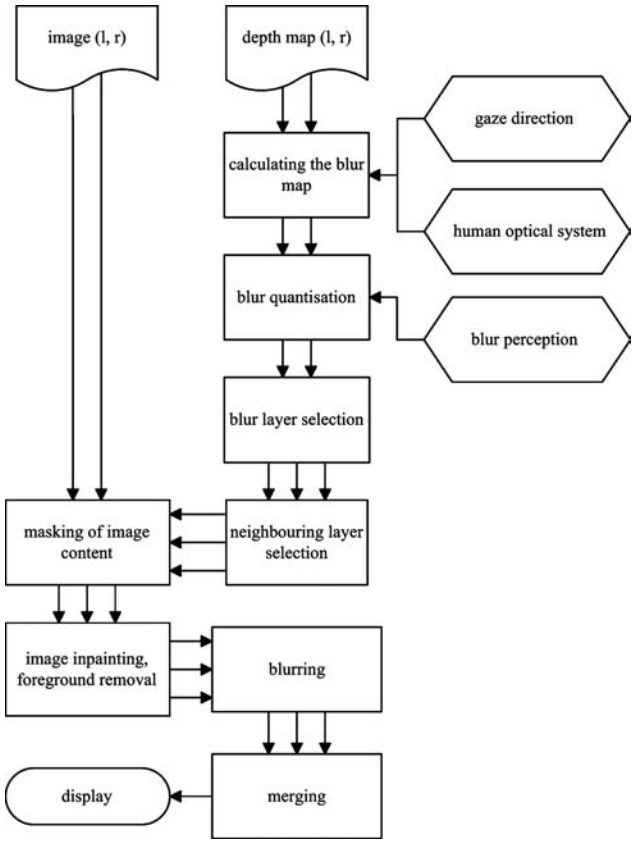

Figure 8 *Block diagram of the algorithm*

for every image point using (5). For accurate blurring, the diameter of the circle of confusion has to be converted in pixel units under consideration of the observer's viewing distance to the display

$$u_{\text{display}} = u_{\text{eye}} \frac{d_{\text{display}}}{i_{\text{eye}}} r_{\text{display}}, \tag{9}$$

where u_{eye} is the diameter of the circle of confusion in the eye, r_{display} is the display resolution and d_{display} is the viewing distance. Every depth value in the depth map is converted in a blur value and we call the resulting map the blur map. We showed in Jaeger [9] that the blur map resolution can be reduced based on the blur perception. This quantisation of the blurriness reduces the computing time significantly. Based on the quantised blur map the image is blurred from the back layer to the front layer. For each layer we remove all foreground objects using image inpainting algorithms. This is necessary owing to the occlusion problem described in Barsky *et al.* [10], where blurred foreground objects have a translucent border resulting in a visible background, which is not included in the source image. Afterwards we blur the inpainted layer with an axially symmetric mean filter with a diameter equal to the circle of confusion diameter. This processing is done for all blur layers. In the final step we merge the blurred layers considering the translucent borders of the foreground objects. This processing has to be done for both stereo images.

Demonstrator

We implemented the described system in a demonstrator using a Zalman TRIMON 3D-Display for stereoscopic presentation. The images for the left and right eye are coded using a left and right circular polariser. Glasses with corresponding polarisers separate the images. The allocation to the left and right eye is done line by line. The gaze estimation is completely separated from the display part. It passes the gaze coordinates as pixel coordinates to the display part. Therefore the system can easily be adapted to other display technologies. We developed a software that reads the depth information from the depth map and presents the blurred and interleaved images for the two eyes. The current system works for a single user. To account for multiple simultaneous users a multi user stereoscopic display has to be used, which tracks the users and creates independent views with independent blurring for them.

Conclusions

We presented a system for gaze direction dependent field depth reduction. The system consists of a stereoscopic display and a gaze tracker. Based on the gaze direction and the viewing distance the field depth is limited either to the natural field depth of the eye or to a content producer predefined field depth. The observer is neither confronted with an image completely in focus, nor restricted to the image areas in focus as in present stereoscopic presentations. First experiences with the system show that the 3D viewing experience is significantly increased using the system. The gaze tracker can be modified to use only one high resolution camera instead of the two cameras. This simplifies the system but decreases the gaze direction estimation accuracy. As long as the accuracy is high enough to identify the gazed image object this is still acceptable. In further research we will analyse the consequences for the viewing experience in detail and the possible extension to a multi user system.

Acknowledgments

The author would like to thank his colleagues for their contributions to this work. He would also like to thank the International Broadcasting Convention for permission to publish this paper.

References

[1] UMBLE E.A.: 'Making it real: the future of stereoscopic 3D film technology', *Computer Graphics Quarterly*, 2008, **40**, (1)

[2] MAIER F.: '3D-Grundlagen, Teil 1', *Professional Production*, 2008, **07 + 08**, pp. 14–18

[3] SCHMIDY R.F., THEWS G.: 'Physiologie des Menschen' (Springer, 1993)

[4] NAHLAOUI M.Y.A.: 'Eye-Tracking - Visuelle blickrichtungserfassung im dreidimensionalen raum', *tm - Technisches Messen.*, 2008, **75**, (7–8), pp. 437–444

[5] POTMESIL M., CHAKRAVARTY I.: 'Synthetic image generation with a lens and aperture camera model', *ACM Transactions on Graphics*, 1982, **1**, (2), pp. 85–108

[6] JAGER T.: 'Nachträgliche Generierung tiefenabhängiger Unschärfe bei elektronischer Bildaufnahme', *FKT*, 2006, **1–2**, pp. 35–41

[7] ATZPADIN N., KAUFF P., SCHREER O.: 'Stereo analysis by hybrid recursive matching for real-time immersive video conferencing', *IEEE Transactions on Circuits and Systems for Video Technology*, 2004, **14**, pp. 321–334

[8] KAUFF P., ATZPADIN N., FEHN C., MULLER M., SCHREER O., SMOLIC A., TANGER R.: 'Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability', *Signal Processing-Image Communication*, 2007, **22**, pp. 217–234

[9] JAGER T.: 'Subjektive Bewertung von Schärfe und Unschärfe in Bildern', *23. Fachtagung der FKTG*, 2008

[10] BARSKY B.A., TOBIAS M.J., CHU D.P., HORN D.R.: 'Elimination of artifacts due to occlusion and discretization problems in image space blurring techniques', *Graphical Models*, 2005, **67**, pp. 584–599

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

User assignment for minimum rate requirements in OFDM-MIMO broadcast systems

C. Huppert F. Knabe J.G. Klotz

Institute of TAIT, Ulm University, Germany
E-mail: carolin.huppert@uni-ulm.de

Abstract: A resource allocation strategy of low computational complexity is proposed that aims at maximising the sum rate while providing minimum rates in a MIMO-OFDM broadcast system. Therefore, two users per carrier are assigned and all eigenvalues of the scheduled users are served by means of beamforming techniques. For each carrier the power distribution between the two users is adapted finally. The algorithm is evaluated by means of simulation results.

1 Introduction

In future communication systems the demands on data rates will increase. To deliver high data rates, techniques such as orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO) are applied. To exploit the resources efficiently sophisticated allocation strategies have to be applied. The optimum solutions are known for many systems; however, they require a high computational complexity owing to their iterative nature. Furthermore, the signalling overhead that is necessary to inform the users about the allocation is usually high and of variable size. To overcome these disadvantages an algorithm that aims at maximising the sum rate under minimum rate requirements was developed in [1] for multi-antenna OFDM broadcast systems where each user has one single receive antenna. This algorithm requires only a small computational complexity compared to the optimum solution presented, e.g. in [2], and the signalling overhead is kept low by serving only two users per carrier. This algorithm applies beamforming techniques such that the transmissions are in the direction of the eigenvalues of the channel. Clearly, this algorithm can also be applied to a MIMO system where the individual users have several receive antennas. This can be done by choosing only the best eigenvalue of each user and proceeding as described in [1]. However, in a system with more than one receive antenna per user the performance of a resource allocation algorithm can be increased by using more than two eigenvalues per carrier while maintaining the restriction of only two users. Thus, an algorithm for minimum rate requirements is derived in this work that takes all eigenvalues of the assigned users into account but still restricts the number of allocated users per carrier to two. To keep the signalling overhead small the power distribution to the eigenvalues of a single user is done uniformly.

2 System model

In this Letter, we consider the downlink of an OFDM-MIMO system with L orthogonal carriers, K users, I transmit antennas at the base station and J receive antennas at each user. The fading coefficient of user k in carrier l from antenna i to receive antenna j is denoted by $H_{k,l}^{(i,j)}$. In addition to the fading we assume Gaussian noise of power N_k for user k. As described above we propose an algorithm that aims at maximising the sum rate while the single rates r_k of each user k are at least equal to the required minimum rate R_M, $r_k \leq R_M$, $\forall k$. To calculate the achievable rates we determine the channel gain matrices $\boldsymbol{G}_{k,l}$ for each user k in each carrier l of size $I \times J$ containing the elements $H_{k,l}^{(i,j)}/\sqrt{N_k}$. Thus, the achievable rate of user k in carrier l can be calculated by the channel matrix $\boldsymbol{G}_{k,l}$ and normalised noise of power $N = 1$. We assume that the channel matrices are of full rank S, i.e. $S = \min\{I, J\}$. We use a

uniform power distribution over the carriers in order to keep the signalling overhead small, $P_1 = P_2 = \cdots = P_L$. Furthermore, we normalise the power for the sake of clarity such that $P_l = 1$.

To use eigen beamforming techniques the channel matrices are decomposed using the singular value decomposition

$$\boldsymbol{G}_{k,l} = \boldsymbol{U}_{k,l} \sum_{k,l} \boldsymbol{V}_{k,l}^{H} \tag{1}$$

where $\boldsymbol{U}_{k,l}$ and $\boldsymbol{V}_{k,l}^{H}$ are unitary matrices and matrix $\sum_{k,l}$ is a diagonal matrix containing the square roots of the eigenvalues as elements.

The achievable rates of two users k_1 and k_2 that are served in carrier l by applying beamforming techniques while serving all eigenvalues with equal power are given by

$$r_{k_1,l} = \log_2 \left| \boldsymbol{I} + \frac{\alpha_l}{S} \sum_{k_1,l} \sum_{k_1,l}^{H} \right| \tag{2}$$

and

$$r_{k_2,l} = \log_2 \left| \boldsymbol{I} + \frac{(1-\alpha_l)/S \sum_{k_2,l} \sum_{k_2,l}^{H}}{\alpha_l/S \sum_{k_2,l} \sum_{k_2,l}^{H} + \boldsymbol{I}} \right| \tag{3}$$

The parameter α_l gives the power fraction that is assigned to user k_1 in carrier l.

3 Resource allocation algorithm

The proposed MIMO algorithm mainly works in three steps. In the first step it aims to maximise the sum rate and the second step is designed to come up to the minimum rates. The third step is performed in this algorithm in order to adapt the power distribution α_l inside the carriers in order to meet the optimisation aim more accurately.

In the following the procedure of this algorithm that is also given in terms of pseudo code in algorithm 1 (Table 1) is described in more detail. To perform the first step, the equivalent fading values $g_{k,l}^{equ}$ have to be determined for each user in each carrier by

$$|g_{k,l}^{equ}|^2 = \left| \boldsymbol{I} + \frac{1}{S} \sum_{k,l} \sum_{k,l}^{H} \right| - 1 \tag{4}$$

These values give the effective fading of a single antenna channel of equivalent rate to the MIMO channel with $\boldsymbol{G}_{k,l}$.

As stated above the first step assigns one user per carrier such that the sum rate is maximised. Therefore each carrier is assigned to its best user where the best user corresponds to the user with the largest $|g_{k,l}^{equ}|$. Since a second user should be assigned to the carriers in the subsequent steps

Table 1

Algorithm 1: Resource allocation algorithm

INITIALISATION
Initialise user rates: $r_k = 0$; $k = 1, \ldots, K$

1. STEP
For all k, l determine $g_{k,l}^{equ}$ according to (4)
for $l = 1 \ldots L$ **do**

- Determine best user k_l in carrier l
- $k_l = \arg\left\{ \max_k \left\{ |g_{k,l}^{equ}| \right\} \right\}$
- and its corresponding rate
- $r_{kl} = \frac{1}{2}\log_2(1 + |g_{k_l,l}^{equ}|^2)$
- and power fraction α_l according to (5)

2. STEP
Determine second users b_l
Create set with indices of carriers with only one user
$\mathcal{U} = \{1, \ldots, L\}$
while $\mathcal{U} \neq \varnothing$ **do**

- Choose user b_l with instantaneous smallest rate,
- $b_l = \arg\{\min_k \{r_k\}\}$
- Choose best carrier c from $\mathcal{U}$
- $c = \arg\{\max_{l \in \mathcal{U}} \{|g_{b_l,l}^{equ}|\}\}$
- Update rate r_{b_l} according to (3)
- Update set of not assigned carriers: $\mathcal{U} = \mathcal{U} \backslash c$

3. STEP
for $k = 1 \ldots K$ **do**

- Determine rate overrun: $O_k = r_k - R_M$
- Determine set $\mathcal{D}_k$ of carriers with k as second user
- **for** $l \in \mathcal{D}_k$ **do**
  - Determine α_l from
  - $\log_2 \left| \boldsymbol{I} + \frac{(1-\alpha_l)/S \sum_{b_l,l} \sum_{b_l,l}^{H}}{\alpha_l/S \sum_{b_l,l} \sum_{b_l,l}^{H} + \boldsymbol{I}} \right| = r_{k,l} - \frac{O_k}{|\mathcal{D}_k|}$
  - $\alpha_l = \min[\alpha_l, 1]$
  - **if** $\alpha_l < 0$ **then**
  - Declare outage; stop algorithm
  - Update rates of both users k_l, b_l in carrier l
  - according to (2) and (3)
  - with the updated power distribution α_l

the power fraction α_l is required to determine the assigned rates. It turns out that it is a good choice to choose the power fraction α_l such that the user k_l assigned to carrier l gets half of its maximum achievable rate, i.e.

$$\log_2 \left| \boldsymbol{I} + \alpha_l \frac{1}{S} \sum_{k,l} \sum_{k,l}^{H} \right| = \frac{1}{2}\log_2(1 + |g_{k_l,l}^{equ}|^2) \tag{5}$$

has to be solved for α_l.

In the second step we try to meet the minimum rate requirements by applying a simple allocation algorithm. Thus, we search for the user $k_{\min}$ with the instantaneous minimum rate and assign it as second user to its best carrier, i.e. the carrier with a maximum $|g_{k,l}^{equ}|$, with only one assigned user yet. The user rate of $k_{\min}$ is updated according to (3).

The third step is used to adapt the power distributions α_l. Concerning the sum rate it is usually advantageous if only the first users are assigned, i.e. choosing $\alpha_l = 1$. On the other hand, even by applying the second step the minimum rate requirements may still not be fulfilled. To consider both these criteria, a third step is performed that should adjust the power fractions α_l such that the first users get more power in carriers where the second users have an instantaneous larger rate than required whereas α_l should be shifted in favour of the second user in carriers where this is necessary to prevent an outage. To cope with these two opposing issues we determine the rate overrun $O_k = r_k - R_M$ for each user. This value is negative for all users that do not fulfil the requirements. In the last step our algorithm tries to minimise the absolute value of this overrun for each user k by adjusting the rates in carriers in which the user k is the second user. Therefore, first a set $\mathcal{D}_k$ is created for each user where the carrier indices with user k as second user are stored. The rate overrun elimination should be distributed uniformly over this set. Thus, for each carrier $l \in \mathcal{D}_k$ the power parameter α_l should be adapted such that the new user rate is

$$\tilde{r}_{k,l} = r_{k,l} - \frac{O_k}{|\mathcal{D}_k|} \tag{6}$$

If this yields a power fraction $\alpha_l > 1$ we just choose $\alpha_l = 1$, whereas an outage is declared for the case $\alpha_l < 0$.

4 Simulation results

In the following the proposed resource allocation algorithm, referred to as *BC MIMO*, is evaluated by means of simulation results. We use a simulation setup with $L = 64$ carriers, $K = 20$ users, $I = 4$ antennas at the base station and $J = 4$ receive antennas at the users. We use independent Rayleigh fading coefficients $H_{k,l}^{(i,j)}$ and choose a stochastic distribution of the noise powers N_k such that the average, normalised signal-to-noise ratios $1/N_K$ are uniformly distributed between 0 and 20 dB. For the simulations we vary the required minimum rate per user R_M. During all simulations no outages occurred.

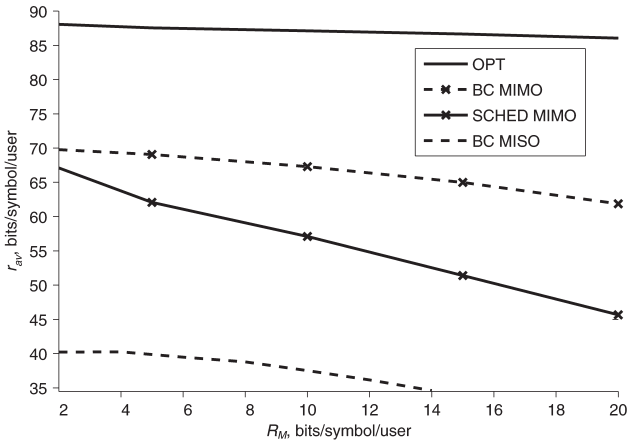

Figure 1 *Achievable rates for different algorithms*

Our algorithm is compared to a scheduling algorithm, *SCHED MIMO*, that searches for the user with the instantaneous minimum rate and assigns it to its best carrier out of those carriers not already used. This is done until the minimum rate requirement is fulfilled for each user. Then the remaining unused carriers are assigned to their best users. Furthermore, the optimal algorithm, *OPT*, and the algorithm from [1], *BC MISO*, as described above are used as comparison.

Fig. 1 shows the average achievable rates per user r_{av}. We see that the proposed algorithm clearly outperforms the *SCHED MIMO* as well as the *BC MISO* algorithm. Comparing the performance of our algorithm with that of the optimum allocation it can be stated that our low complex algorithm has a performance degradation of less than 30% in the considered range.

5 Conclusion

We propose an allocation algorithm with two users per carrier for an OFDM-MIMO broadcast system with minimum rate requirements. By means of simulation results we show that this algorithm is superior to an allocation with only one user per carrier in terms of the achievable rate. Compared to the optimum allocation our proposed strategy of low complexity performs in a reasonable range.

6 Acknowledgment

The work described in this Letter was supported by the German research council 'Deutsche Forschungsgemeinschaft' (DFG) under Grant No. 867/18-1 and Grant No. 867/19-1.

7 References

[1] HUPPERT C., NEUMANN D., WECKERLE M., BOSSERT M.: 'Resource allocation with minimum rates for MISO-OFDM broadcast channels'. Proc. 7th Int. ITG Conf. on Source and Channel Coding, Ulm, Germany, January 2008

[2] WUNDER G., MICHEL T.: 'Minimum rates scheduling for MIMO-OFDM broadcast channels'. Proc. 9th IEEE Int. Symp. on Spread Spectrum Techniques and Applications (ISSSTA 2006), Manaus, Brazil, August 2006

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

Fuzzy logic congestion control for broadband wireless IPTV

E.A. Jammeh M. Fleury M. Ghanbari

Department of Computing and Electronic Systems, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
E-mail: fleum@essex.ac.uk

Abstract: It is demonstrated that interval type-2 fuzzy logic control (IT2 FLC) is more robust than traditional fuzzy logic congestion control of video streaming. An IT2 FLC is compared to the well-known TFRC and TEAR congestion controllers for Internet multimedia streaming. On an all-IP network with broadband wireless access, delivered video quality is improved with the IT2 FLC by about 1 dB for each client, once the offered traffic (up to 50 video streams) exceeds the capacity available to video over the wireless link.

1 Introduction

All-IP broadband networks are being created with multimedia bandwidth requirements in mind. A unicast IPTV service forming a pipe or sub-channel on the converged network may need to negotiate a broadband wireless link. Where there is a need for multiple variable bit rate video streams to share the same pipe a problem of link utilisation arises [1], requiring congestion control at the server bank. Conventional controllers such as TCP-friendly rate control (TFRC) [2] and TCP emulation at receivers (TEAR) [3], originating in a TCP-dominated internet, will stream video up to the capacity of the pipe, but reacting to feedback may overestimate the capacity, resulting in packet loss, which leads to reduced video quality. In this Letter, fuzzy logic control (FLC) is shown to outperform conventional control in such a network by changing the quantisation parameter for live video or through a bit rate transcoder for pre-encoded video. Moreover, compared to prior use of traditional (type-1) fuzzy logic for similar purposes [4], interval type-2 (IT2) FLC has been employed, as this is robust to feedback measurement uncertainties.

2 Methodology

In Fig. 1, FLC units independently control the bit rate of a number of bitstreams destined for the broadband wireless channel. Through FLC either the quantisation parameter at a codec is directly changed or the bit rate is indirectly altered by a transcoder. In tests, the broadband wireless link is modelled as the bottleneck link in a 'dumbbell' topology network.

In [5], a type-1 FLC was demonstrated that based its response on measured packet delay at a receiver. However, traditional FLC is not completely fuzzy, as the boundaries of its membership functions are fixed. IT2 FLC [6] can address this problem by extending a footprint of uncertainty (FOU) on either side of an existing type-1 membership function. In IT2 fuzzy logic, the variation is assumed to be constant across the FOU, hence the designation 'interval'. By restricting variation of the secondary membership functions within the FOU, IT2 decisions can be computed in real-time, which would not be the case if general T2 logic without restriction were to be used.

In preliminary comparison tests, between the original FLC [5] and an IT2 version, the well-known ns-2 network simulator (v. 2.32) was used, with type-1 and IT2 FLC implemented as new protocols within ns-2. Video was streamed across the bottleneck with reduced capacity of 0.4 Mbit/s and 5 ms delay. Additive Gaussian noise was superimposed upon the delay measurements passed to both controllers. A normal distribution generated a random noise value with zero mean and a specified standard deviation, determined by the level of noise required and dynamically adjusted relative to the measured (simulated) value. For each simulation the level of additional noise was

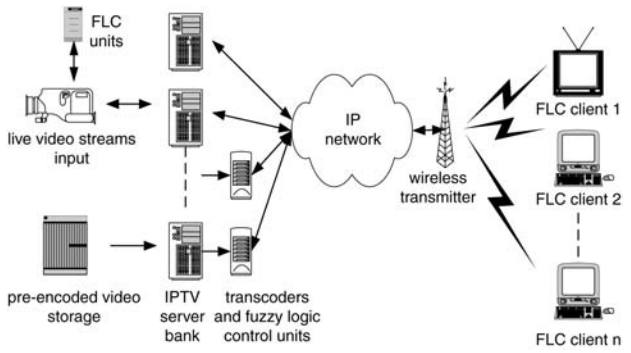

Figure 1 *IPTV video delivery architecture*

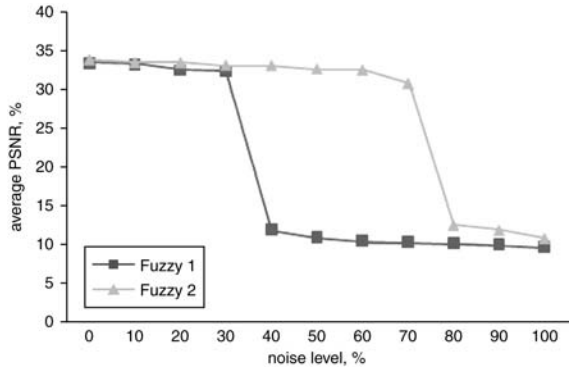

Figure 2 *Mean received video quality (Y-PSNR) for an increasing noise level*

incrementally increased. At each incremental step, the relative performance was judged in terms of delivered video quality (PSNR). Input was a 40s MPEG-2 encoded video clip, showing a newsreader with a changing backdrop, with moderate movement. The variable bit rate (VBR) 25 frame/s SIF-sized clip having a GOP structure of N = 12, M = 3 was encoded with a mean bit rate of 1 Mbit/s. For error resilience purposes, there was one slice per packet, resulting in 18 packets per frame. The FLC controllers adjusted their rate every frame based on an averaged measure of packet delay, which occurs through queuing at the input router.

The IT2 FLC was then compared to TFRC [2] and TEAR [3], using the latter's publicly-available ns-2 models. Without a transcoder, TFRC and TEAR require playout buffers to smooth out network delay. Therefore, PSNR is affected by loss rate only, assuming a large enough buffer to avoid overflow. FLC also reduces the video quality through transcoding if there is insufficient bandwidth, but this avoids the need for long start-up delays and allows smaller buffers on mobile devices. The particular transcoder modelled has a cost in that the sending rate can be no more than 90% of the original. Both TFRC and TEAR rely on measurements of the round trip time (RTT), though packet loss also plays its part.

For these comparisons the pipe's capacity across the broadband link was set to 25 Mbit/s. The one-way delay, modelling the latency across the complete network path, was now set to 40 ms, which is the same as the maximum delay across a country such as the UK. As in the type-1 and IT2 comparison, side link delay was set to 1 ms and the side link capacity was set at 100 Mbit/s to easily cope with the input video rate. The mean encoded video rate was again 1 Mbit/s. Again, the buffer size on the intermediate routers was set to RTT × bandwidth, to avoid overflow through too small a buffer. The router queuing discipline was drop-tail. The intention of these tests was to see how many video streams could be accommodated across the bottleneck link and consequently the channel was kept error free.

The number of controlled video sources (replicating the news clip source) was incrementally increased. The starting times of streaming the 'news clip' to each client was staggered, and then each clip was repeatedly sent over 200 ms. The first 40 s of results were discarded as representing transient results. This method was chosen, rather than selecting from different video clips, because the side effects of the video clip type do not intrude.

3 Results

In the preliminary comparison, Fig. 2 shows that beyond 30% additional noise, the IT2 FLC congestion controller (upper plot) achieved significant improvement over the type-1 FLC in terms of better average video quality. For very high levels of additional noise, the quality is very poor whatever the controller. Furthermore, Table 1 demonstrates the superiority of IT2 FLC to type-1 FLC, when sending the same video stream, through reductions in the standard deviation of the sending rates, which implies that the delivered video quality will fluctuate less. Therefore,

Table 1 Standard deviations of FLC type-1 and type-2 sending rates

Noise level (%)	Type-1 (kbit/s)	Type-2 (kbit/s)
0	77.53	76.72
10	78.19	76.61
20	78.99	77.10
30	80.28	77.68
40	109.93	77.75
50	193.61	78.24
60	227.17	80.24
70	230.02	84.29
80	230.65	93.82
90	230.92	113.36
100	231.08	124.65

Table 2 Comparison of IT2 FLC and other congestion controllers, for an increasing number of video stream sources

No. of sources	No control			TFRC		
	Loss rate (%)	Link use (%)	PSNR (dB)	Loss rate (%)	Link use (%)	PSNR (dB)
10	0.0	40.0	–	0.47	100.44	38.66
15	0.0	60.0	–	0.82	100.80	38.33
20	0.0	80.0	–	1.17	101.16	37.31
25	0.0	100.0	–	1.50	101.48	36.08
30	16.66	120.0	–	1.81	101.80	35.11
35	28.56	140.0	–	2.11	102.80	33.78
40	37.49	160.0	–	2.39	102.44	33.07
45	44.44	180.0	–	2.65	107.78	31.34
50	49.99	200.0	–	2.91	102.96	30.18
	IT2 FLC			TEAR		
10	0.0	35.92	38.92	0.47	100.40	36.75
15	0.0	53.88	38.87	0.95	100.88	36.23
20	0.0	71.84	38.86	1.62	101.56	34.65
25	0.0	89.92	38.12	2.50	102.52	33.27
30	0.0016	99.96	37.10	3.51	103.60	32.34
35	0.0026	99.96	36.10	4.61	104.80	31.56
40	0.0029	99.96	34.70	5.75	106.08	30.70
45	0.0038	99.84	32.49	6.86	107.36	29.61
50	0.0048	99.82	30.73	7.91	108.56	28.78

Table 1 shows that the subjective experience with IT2 FLC would be better even if the average objective quality was similar to that of type-1 FLC, because of the reduced level of fluctuations.

Turning to the comparison with non-FLC, in Table 2 when there is no control, there is no packet loss until the capacity of the link is reached. Thereafter, the link utilisation grows and, as might be expected, packet loss rate rapidly climbs. Failure to estimate the available bandwidth causes both TFRC's and TEAR's mean link use in Table 2 to exceed the capacity of the bottleneck link. As the number of flows increases, it becomes increasingly difficult to control the flows and there is a steady upward trend in the overshoot. In particular for TEAR, this leads to considerable packet loss. The packet loss for 'no control' becomes so high that the PSNR does not stand comparison in Table 2, and consequently is not recorded. Because the IT2 FLC sending rates are constrained by the bit rate transcoders, when the offered video streams' rate is below the 25 Mbit/s capacity the link use is lower than that of TFRC and TEAR. However, for offered rates exceeding the capacity, FLC avoids overshoots resulting in efficient utilisation and improved delivered video quality over the better of the two traditional controllers, TFRC, by around 1 dB.

4 Conclusion

IT2 FLC is more robust than a type-1 controller, bringing confidence in its ability to adapt to adverse network conditions. Perhaps surprisingly, TFRC outperforms the later TEAR. However, FLC is not only more suited to streaming IPTV in a broadband network than either of these controllers but will also result in improved user satisfaction.

5 Acknowledgments

This work was supported by the EPSRC, UK, under grant no. EP/C538692/1. The authors acknowledge assistance from C. Wagner and H. Hagras in applying interval type-2 fuzzy logic to the authors' original type-1 controller.

6 References

[1] MUNTEAN G.M.: 'Efficient delivery of multimedia streams over broadband networks using QOAS', *IEEE Trans. Broadcast.*, 2007, **52**, (2), pp. 230–235

[2] HANDLEY M., FLOYD S., PADYHE J., WIDMER J.: 'TCP friendly rate control (TFRC): Protocol specification'. IETF RFC 3448, 2003

[3] RHEE I., OZDEMIR V., YI Y.: 'TEAR: TCP emulation at receivers – flow control for multimedia streaming'. NCSU Technical Report, April 2000

[4] REZAEI M., HANUKSE M.H., GABBOUJ M.: 'Semi-fuzzy rate controller for variable bit rate video', *IEEE Trans. Circuits Syst. Video Technol.*, 2008, **18**, (5), pp. 633–645

[5] JAMMEH E.A., FLEURY M., GHANBARI M.: 'Delay-based congestion avoidance for video communication with fuzzy logic control'. Int. Packet Video Workshop, Lisbon, Portugal, November 2007

[6] MENDEL J.M.: 'Type-2 fuzzy sets and systems: an overview', *IEEE Comput. Intell.*, 2007, **2**, (1), pp. 20–29

Papers and articles on electronic media technology from IBC 2009 presented with selected papers from the IET's flagship publication *Electronics Letters*

ISSN 2041-5923

3D motion estimation for depth information compression in 3D-TV applications

B. Kamolrat W.A.C. Fernando M. Mrak

Centre for Communication System Research, University of Surrey, Guildford GU2 7XH, United Kingdom
E-mail: b.kamolrat@surrey.ac.uk

Abstract: A new approach for predicting and coding depth information in 3D-TV (three-dimensional television) applications is presented. Properties of the depth information, which complements monoscopic video and enables 3D experience, are used in the proposed 3D motion prediction. The new approach leads to more efficient motion compensation and finally to higher compression. Built on the top of the conventional approaches for video coding, the proposed technique is suitable for integration in upcoming 3D-TV products.

1 Introduction

With significant success in both display and network technologies [1, 2], commercial implementations of three-dimensional television (3D-TV) can be realised. Modern approaches recently developed for 3D-TV are based on monoscopic video (simply called colour) and the depth information [3], which is used to synthesise stereoscopic video channels. In video compression, the depth information is treated as a greyscale monoscopic video since it consists of pixels whose values are related to the depth position of the corresponding pixels in the colour image [4]. Even though basic concepts on which the video coding is based, i.e. temporal and spatial predictions, are efficient in compression of colour sequences, they are not efficient for coding depth images. Recently, many techniques have been proposed in a video coding community to improve the coding efficiency in depth sequences. For example, in [5] a novel mesh-based depth coding is introduced to encode depth images. In [6], the concept of region-of-interest (ROI) is applied to ensure a good quality at important parts of depth sequences. The rapid growth of 3D video also draws intension from the MPEG standardisation organisation, and the representation format for colour and depth images was developed under the name of MPEG-C [7]. In such 3D video representation the depth information has its specific inter-frame properties since it accurately reflects movements in the third dimension. To exploit specific video properties and to enable support for new application requirements, different approaches for motion compensation have been recently introduced. For example, in [8], the authors use a flexible motion model to reduce the motion cost for video frames with uniform areas; in [9] flexible optimisation of motion models is used for support of quality scalability. Observing the fact that, unlike in the monoscopic video, the inter-frame changes in the depth map are not addressed by 2D video coding tools, in this Letter a new approach for motion compensation (MC) is proposed. This new approach exploits characteristics of the depth information frames and enables better prediction of frames, leading to enhanced compression.

2 3D motion estimation

In block-based motion compensation the frames are partitioned into blocks. At the encoder the motion estimation (ME) is performed in order to find corresponding blocks in reference frames, from which the blocks in the current frames are predicted. In conventional video coding, which is designed for monoscopic video, efficient compensation is performed using blocks from spatially different areas from the neighbouring frames. In addition to the spatial displacement, here a depth displacement is also considered. To achieve the best matching in such 3D space, new motion estimation is proposed. For a block $\boldsymbol{B}$ in the current frame, the best match is found in the reference frames ($\boldsymbol{F}_i$, where i is the reference frame index). For each reference frame i the motion vector $\boldsymbol{v}_i$ is determined in ME such that the sum of squared errors between the current block and the block

from the reference frame is minimised, as follows:

$$v_i = \underset{v=(x.y.z)}{\operatorname{argmin}} \sum_{a=1}^{N}\sum_{b=1}^{N}(B(a,b) - (F_i(a+x, b+y)+z))^2 \quad (1)$$

where (x, y) are the spatial motion vector components and z is the depth motion vector component. $B(a, b)$ and $F_i(a, b)$ are the values of a pixel at position (a, b) of block-size N^2 pixels in the current block and a block in the reference frame, respectively. To enable searchs for different depth positions, the search space is extended from traditional 2D space to the 3D space. For the proposed 3D ME, the block searching is performed in the depth dimension over the reference frame $\boldsymbol{F}_i$ and the search spaces are

$$\{\boldsymbol{F}_i^{-K}, \boldsymbol{F}_i^{-K+1}, \ldots, \boldsymbol{F}_i^{0}, \ldots, \boldsymbol{F}_i^{K}\} \quad (2)$$

where $\boldsymbol{F}_i^k$ is a version of the reference frame, the values of which are shifted for a value k, i.e. for the depth shift. Note that the 3D search point with $k = 0$ corresponds to the 2D search. In this approach the depth component (z) of a 3D motion vector is an integer value in the range $(-K$ to $K)$ and this range is a searching window in the depth dimension. An example of the depth frames, the pixel values of which are varied during the depth searching process, is shown in Fig. 1. For the reference frame shown in Fig. 1*b*, by adding values of $k < 0$, objects within the scene are shifted away from the camera (Fig. 1*a*) leading to a darker colour. On the other hand, if values of $k > 0$ are added to the reference frame, the objects are shifted closer to the camera (Fig. 1*c*) leading to a brighter colour.

3 3D video coding scheme

As in the traditional video coding schemes, the proposed depth information coding also used both unidirectional prediction (P frames) and bidirectional prediction (B) frames. For P frames only forward prediction is used, while for B frames a mixed mode of forward and backward modes is used, i.e. selection of the mode is based on minimisation of a prediction error. All motion vector components are encoded in a predictive way using values of neighbouring, already-encoded vectors. The difference is encoded with a lossless entropy coder that uses variable length codes.

Figure 1 *Depth information for reference frame original (Fig. 1b) and frames with lower (Fig. 1a) and higher (Fig. 1c) pixel values for 3D MF*

a $\mathbf{F}_i^{-K}$
b $\mathbf{F}_i^{0}$
c $\mathbf{F}_i^{K}$

4 Selected experimental results

Three popular test sequences were used in the experiments. These sequences represent a variety of depth changes, motion activity and complexity. All sequences are in CIF (352 × 288) resolution at a frame-rate of 25 fps. For both approaches the group of picture (GOP) size is 15 with IBBPBBP... structure. Each frame is partitioned into blocks of $n \times n$ pixels. Only the compression performance of the depth information is reported since the new method does not influence the performance of compression of the colour channel.

The efficiency of the proposed method is demonstrated at an example of a depth information frame that is shown in Fig. 2*a*. For a selected area of motion blocks (size 16 × 16 pixels), which has large depth change compared to the reference frame, the motion compensated signals after performing 2D MC and 3D MC are represented in Figs. 2*b* and *c*, respectively. Black regions correspond to efficiently compensated areas. Brighter areas correspond to higher compensation errors. It can be observed that at the selected area, the proposed 3D MC provides better motion compensation, which then improves compression efficiency.

In addition to improved compensation, an important factor of overall 3D video codec is the bit-rate. The overall bit-rate consists of the residual rate (decreased because of highly efficient compensation) and the motion bit-rate. With the proposed method the motion bit-rate increases as demonstrated in Table 1 because a set of additional MVs (third dimension) needs to be transmitted. The presented motion bit-rates are averaged over all test points from

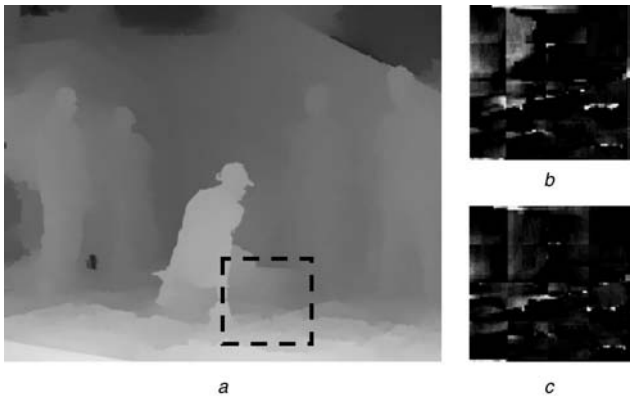

Figure 2 *Motion compensated frame after performing 2D MC and 3D MC*

a First P frame from test sequence 'Break-dance' with selected area with large depth change
b Selected blocks from motion compensated frame for 2D MC for area with large depth change
c Selected blocks from motion compensated frame for 3D MC for area with large depth change

Table 1 Averaged bit rates used to encode motion information of depth sequences for 2D MC and 3D MC in kbit/s

Video sequence/ MC method	'Orbi'	'Ballet'	'Break-dance'
2D MC	105.78	134.17	157.15
3D MC	149.94	198.54	217.69

Fig. 3. However, this is compensated for by high reduction in the residual bit-rate, which then results in significant overall bit-rate reduction as demonstrated by the final results, as follows. Fig. 3 represents final rate-distortion performance for the proposed 3D video coding solution. At high bit-rates, where the quality is also high, the proposed method (3D MC) always achieves significantly better results than the traditional motion compensation based on 2D prediction (2D MC). This is because the large savings introduced by more efficient prediction (less bits required for coding residual) compensates for the increase of motion bit-rate owing to an additional MV in the third dimension. The gain is largest for sequences with rapid depth changes, such as for the 'Break-dance' test sequence. For the sequences that have minor depth changes (e.g. the 'Orbi' sequence), there is not much gain with 3D MC. At low bit-rates the increased motion information can become significant and therefore the 2D MC can outperform 3D MC. However, at those operating points the quality is poor, which is not considered in 3D-TV broadcasting, thus it is recommended to use resolution reduction in order to maximise low bit-rate performance. For broadcasting applications the 3D MC brings significant gains while keeping the decoder complexity low and is therefore suitable for adoption in future 3D video products.

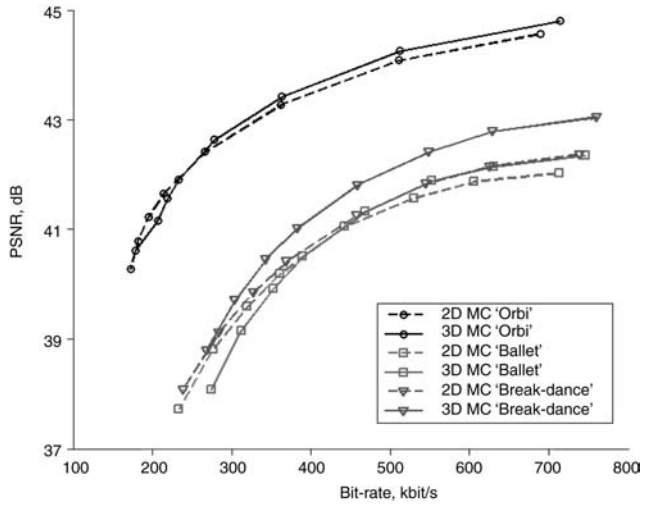

Figure 3 *Decoding results for depth map decoding using 2D MC and 3D MC*

5 Conclusions

The results show that the proposed application of 3D prediction in 3D-TV applications leads to enhanced compression of the depth information, compared to the traditional 2D compensation. Although a full 3D search adds computational complexity to the encoder, the decoding complexity remains almost equal to the complexity of the 2D case because only simple additions are needed to the depth values. Experimental results show that the proposed method is capable of significantly increasing video quality at given 3D-TV operating points.

6 References

[1] Phillips Research Press Information, 'Phillips 3D Information Display Solution Adds Extra Dimension to in-store messaging', http://www.research.philips.com/newscenter/archive/, September 2005

[2] SON J.-Y., JAVIDI B., KWACK K.-D.: 'Methods for displaying three-dimensional images', *Proc. IEEE*, 2006, **94**, pp. 502–523

[3] A. BOURGE AND C. FEHN: 'ISO/IEC CD 23002-3 Auxiliary Video Data Representation,' ISO/IEC JTC 1/SC 29/WG 11/N8038, 2006

[4] FEHN C.: 'Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV'. Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI, January 2004, pp. 93–104

[5] KIM S.-Y., HO Y.: 'Mesh-based depth coding for 3D video using hierarchical decomposition of depth maps'. IEEE Int. Conf. on Image Processing, September 2007, pp. 117–120

[6] KARISSON L.S., SJOSTROM M.: 'Region-of-interest 3D video coding based on depth images'. 3DTV Conf.: The True Vision-Capture, Transmission and Display of 3D Video, May 2008, pp. 141–144

[7] ISO/IEC 23003-3, 'Mpeg-c part 3: Representation of Auxiliary Video and Supplemental Information,' October 2007

[8] KAMOLRAT B., MRAK M., FERNANDO W.A.C.: 'Flexible motion model with variable size blocks for depth frames coding in colour-depth based 3D Video Coding'. IEEE Int. Conf. on Multimedia & Expo, June 2008, pp. 573–576

[9] MRAK M., SPRLJAN N., IZQUIERDO E.: 'Motion estimation in temporal subbands for quality scalable motion coding', *Electron. Lett.*, 2005, **41**, (19), pp. 1050–1051