IET Journals
The Institution of
Engineering and Technology

Vol 3 | ISSN 2041-5923
2011

The Best of IET and IBC

INSIDE Papers and articles on electronic media technology from IBC 2011 presented with selected papers from the IET's flagship publication *Electronics Letters*.

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

www.**theiet**.org/journals

Contents

VOLUME 3 2011

Selected content from IBC2011

Selected content from the IET

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

ISSN 2041-5923

Introduction

In this volume

Welcome to The Best of IET and IBC, 2011, the third volume of the joint publication of the International Broadcasting Convention and the Institution of Engineering and Technology. Once again, you will find on these pages the very best technical content from this year's IBC combined with some of the best broadcasting technology research from the last year published in *Electronics Letters*, the flagship publication of IET Journals. As an original member of the IBC partnership board, the IET continues to work closely with the IBC to develop this annual publication to produce something that truly represents the exciting nature of current broadcasting technology research.

Representing the content of IBC2011 this volume contains the seven best papers presented at this year's conference, selected by the IBC Technical Papers Committee in conjunction with the executive board of the IET's Multimedia Communications community. This collection begins with the paper selected as the best in the conference, Nishida *et al.*'s paper on NHK's super hi-vision system. Before that, as part of the IBC's priority for developing future and emerging broadcasting professionals, our feature editorial this year is written by Kristin Mason, leader of the IBC's Rising Stars Programme. Kristin gives us her insight into this fresh initiative, which was piloted at IBC2010 and which aims to help students and other attendees new to the broadcasting industry get the most out of their time at IBC and develop their understanding of the industry.

Rounding off the IBC2011 content, but continuing our focus on the future, we have an interview with Anders Prytz. Anders's paper looking at video quality assessment was selected as this year's best young professional contribution and in his interview Anders provides some more background on both his work and his experiences as a young researcher in this industry and at IBC.

Representing the content of IET Journals we include three papers from *Electronics Letters.* Two of these are from the area of video processing, one looking at a challenge developing with the increasing volume of 3D video media, and one at using sports content across very different screen sizes, an issue introduced by the use of video footage through multiple devices, from big screen TVs to palm-of-the-hand smartphones. Our third representative from *Electronics Letters* is from the maturing arena of intuitive human–machine interfaces, of interest both for home entertainment and interactive advertising. Each Letter is complemented by an interview with the authors.

We are sure you will find this volume an interesting sample of the best of IBC2011 and the high-quality peer-reviewed research published by IET Journals and we would like to extend our thanks to everyone involved in creating the 2011 volume of The Best of IET and IBC. We hope you will enjoy it, and wish all of you attending this year a stimulating IBC2011.

Michael Lumley
Chairman of the IBC Conference
&
The IET Multimedia Communications Network executive team

© The Institution of Engineering and Technology 2011

Who we are

IBC

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

The IET

The IET is one of the world's leading professional societies for the engineering and technology community, with more than 150 000 members in 127 countries and offices in Europe, North America and Asia-Pacific. It is also a publisher whose portfolio includes a suite of 25 internationally renowned peer-reviewed journals covering the entire spectrum of electronic and electrical engineering and technology. Many of the innovative products that find their way into the exhibition halls of IBC will have originated from research published in IET titles with more than a third of the IET's journals covering topics relevant to the IBC community (e.g. *IET: Image Processing*; *Computer Vision*; *Communications*; *Information Security*; *Microwave Antennas & Propagation*; *Optoelectronics*, *Circuits & Systems* and *Signal Processing*). The IET Letters contained in this publication come from the IET's flagship journal, *Electronics Letters*, which embraces all aspects of electronic engineering and technology. *Electronics Letters* has a unique nature, combining a wide interdisciplinary readership with a short paper format and very rapid publication; produced fortnightly, in print and online. Many authors choose to publish their preliminary results in *Electronics Letters* even before presenting their results at conference, because of the journal's reputation for quality and speed. In January 2010 *Electronics Letters* was given a fresh new look, bringing its readers even more information about the research through a contemporary colour section that includes author interviews and feature articles expanding on selected work from each issue.

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

© The Institution of Engineering and Technology 2011

ISSN 2041-5923

Editorial

IBC's Rising Stars Programme

Broadcasting is nothing without the people who make things happen, both creatively and technically, and IBC's Rising Stars Programme is specifically for students and new entrants who will form the broadcasting industry of the future. Alongside the acknowledgment that there is increasingly a shortage of young engineers and broadcasting professionals coming through the system, it is also the case that those entering the broadcasting industry do so with an ever-widening range of expertise and qualifications in subjects and disciplines that would not have been part of broadcasting a few years ago.

With 1300+ exhibitors, in excess of 48,000 visitors and a wealth of information from a packed conference programme, it is easy to see why IBC could be a little daunting to the first-time visitor. When that visitor is also new to the industry, it can be nothing if not overwhelming and the IBC's Rising Stars Programme is specifically designed to enable students and those new to the broadcasting industry to make the most of their time at IBC. The programme therefore provides a supportive environment, plenty of handy hints for IBC survival (comfortable shoes!) and lots of opportunities to get to grips with the industry they are joining.

Piloted last year, the IBC's Rising Stars Programme combines a range of free sessions designed to be distinctive from, and complementary to, the main IBC conference and IBC exhibition. The programme of sessions runs alongside the main conference and 'added value' sessions, and Rising Star delegates are encouraged to tailor their own IBC experience, mixing conference sessions with Rising Star sessions to suit what they need, with attention specifically drawn to conference sessions such as 'What Caught My Eye' and Keynotes that help to set issues in context.

The 2011 Rising Stars Programme features four days of specially-designed sessions with day one, Friday 9th September, setting the scene for IBC itself and the changing media landscape. Saturday and Sunday feature more general sessions including a window on the workings of Aardman Animation as they work in both record-breakingly small and large scale. There will also be the opportunity to hear first-hand about the experiences of transforming a 2D children's TV animation into a live-action 3D feature film. The Monday will include sessions on career planning, working abroad and how to approach it, and entrepreneurship.

Above: This year three presenters from the 2010 Rising Stars Showcase will film interviews with industry figures talking about their career and the future of their area.

Most of the IBC's Rising Stars Programme takes place in a single 'hub' room with lots of opportunity for networking and, in contrast to many of the IBC conference sessions, most are single-speaker in format rather than panel sessions. This allows an easy level of formality and for the audience to learn from established industry figures. Just as in the main conference, the contributors to the student programme have a huge range of experience across the broadcasting industry. All of the sessions have been designed to allow plenty of time for questions and discussion to allow the audience to understand some of the complex issues facing the industry and make the most of having such expertise available.

As part of the IBC's Rising Stars offering, daily tours will guide participants around the IBC exhibition and, as far as is possible, each will be tailored to focus on the requirements of the group, whilst providing an experienced overview of the exhibition as a whole and the trends in today's world of broadcasting.

The IBC's Rising Stars Showcase session is a main 'added value' conference session in which teams of students present on a broadcast-related issue of their choosing. The

© The Institution of Engineering and Technology 2011

session in the 2010 pilot saw four teams present on such diverse subjects as a business model for digital film distribution, Ravensbourne's new Media Asset Management system, and internet privacy from Westminster University. The University of Southern California's presentation on the 'synaptic crowd' was a piece of internet technology that allowed for web-based vox pops, which had everyone in the room discussing the implications for its potential use. This year's Showcase session promises an equally varied range of presentations, so be there or miss out!

With IBC being run by the industry, for the industry, the Rising Stars Programme is for new entrants to the industry, and also involves them in what is offered. Throughout the planning for the IBC's Rising Stars Programme, the ethos has been to build a legacy both in terms of content and taking into account the feedback of previous Rising Stars. So, Craig Gardner, one of the IABM-sponsored IBC delegates from last year has been involved in the planning for this year's Rising Stars Programme and three of those who presented in the Rising Stars Showcase session last year are forming the crew for shooting a number of industry interviews during IBC this year. These specially-shot, short career-based interviews will be available to view online and will add to a growing legacy of content for IBC's Rising Stars. Each short interview will feature an industry figure revealing how they started in the industry, what they do now, and where they predict their area of the industry will go in the next few years. Online, these inspirational interviews will demonstrate the considerable range of roles, backgrounds, and experience across the industry. IBC's Rising Stars presentation material will also be made available through the IET website: www.theiet.org/ibc.

Attending IBC pretty regularly over more decades than should be admitted, I can truthfully say that I have never done so without learning something - often in an area I had not predicted! Sometimes it has been the clear demonstration of an industry trend, other times it has been something, perhaps in the area of new technology that is just unbelievably neat, inspiring, and appropriate for its purpose.

As an industry, we need those already in it and those joining it, to feel part of broadcasting in all its guises. We gain nothing if we are precious about our experience and fail to hand on our expertise. In offering their Rising Stars Programme, IBC has clearly demonstrated that it recognises this and is prepared to invest in the broadcasting industry of the future.

Kristin Mason
Leader, IBC's Rising Stars Programme

© The Institution of Engineering and Technology 2011

Selected content from IBC2011. Winner of the Best Paper award

ISSN 2041-5923

Super hi-vision system offering enhanced sense of presence and new visual experience

Y. Nishida K. Masaoka M. Sugawara K. Ohmura M. Emoto E. Nakasu

NHK (Japan Broadcasting Corporation), 1-10-11 Kinuta Setagaya-Ku, Tokyo 157-8510, Japan
E-mail: nishida.y-fe@nhk.or.jp

Abstract: NHK has invested considerable effort in researching and developing an extremely high-resolution video system called super hi-vision that can provide an increased sense of presence as well as the sense of realness to viewers. It has been designed with the optimum video parameters for two-dimensional motion images on the basis of the psychophysical characteristics of human vision. This paper describes the video parameters of super hi-vision, including spatial and temporal resolutions and tone and colour reproduction, and the importance of super hi-vision as a future television system. Super hi-vision is suitable for various viewing environments, ranging from theatrical to home and mobile environments; it allows the use of displays with different sizes, and different viewing distances. This high-fidelity system provides viewers with a new visual experience in all environments, surpassing the existing television systems.

1 Introduction

The objective of super hi-vision is to provide a viewing experience much better than that provided by existing television systems. Super hi-vision is the ultimate two-dimensional television to have been developed at NHK. We have set 2020 as the target year to start experimental broadcasting with a broadcasting satellite in the 21 GHz band. To achieve this objective, we have invested considerable effort in the research and development of super hi-vision. We have focused on identifying the main video and sound parameters and on developing an end-to-end system from cameras to displays; the end-to-end system involves the consideration of data storage, compression coding, and data transmission.

Video parameters, including spatial resolution, temporal resolution, tone reproduction, and colour representation, characterise the performance of a video system. The following fundamental requirements have been taken into account in determining the parameter values of super hi-vision: worthwhile improvement in quality beyond HDTV; compatibility, interoperability, and commonality with HDTV; and technical feasibility in the foreseeable future. We have conducted a number of human scientific studies to investigate the psychophysical effects of the video parameters and to determine suitable parameter values.

2 Spatial resolution

Super hi-vision has been designed to provide an enhanced sense of presence and a new visual experience. The field-of-view (FOV) angle and angular resolution in terms of the number of pixels per arc-degree of the FOV angle determine the spatial resolution. The sense of presence may involve various aspects. Among them, a 'sense of being there' and a 'sense of realness' have been considered as being the factors that distinguish super hi-vision from existing television systems.

2.1 Sense of being there

A wider FOV leads to a stronger sense of being there. Fig. 1 shows a result of a subjective assessment of the horizontal FOV and the sense of being there [1]. The sense of being there increases with the FOV and saturates at an FOV of around 80–100°. In the experiment, four images shot with a camera angle of 60° were presented to participants

© The Institution of Engineering and Technology 2011

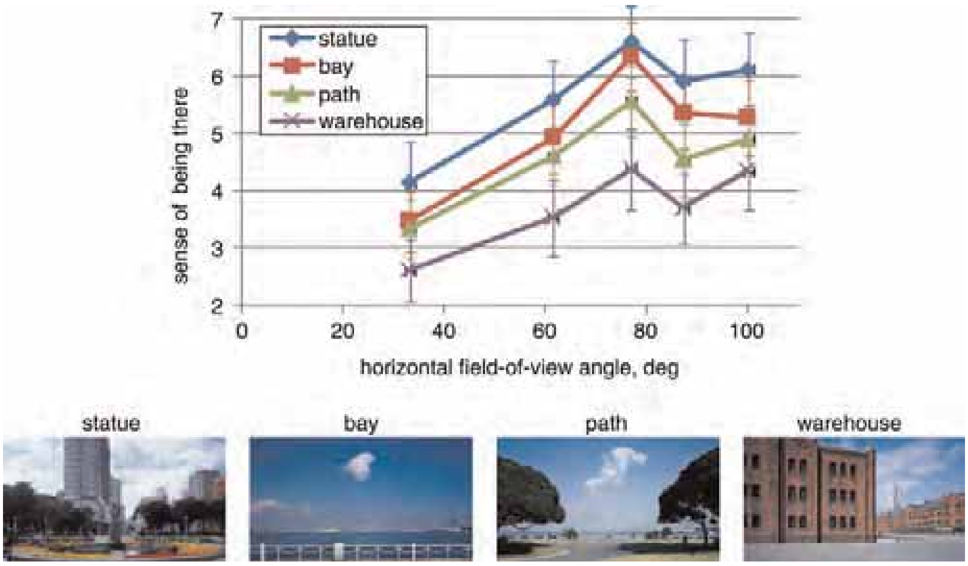

Figure 1 *Horizontal field-of-view angle and sense of 'being there' (mean ± standard error)*

at five different FOV angles. Each participant evaluated the degree of sense of being there from the images on a continuous scale ranging from 0 (none) to 10 (extreme). A total of 200 participants were employed, and they were divided into five groups of 40 participants. Each group performed the evaluation for one of the FOV angles. Another similar experiment performed in the same study by using images obtained with a camera angle of 100° showed a similar result.

This result indicates that the target FOV for super hi-vision should be around 80–100°. This corresponds to a viewing distance that is 0.75 to 1.00 times the picture height (0.75–1 H), at which distance people with normal visual acuity cannot just discern the pixel structure.

2.2 Sense of realness

The spatial resolution is responsible for the sense of realness or visual fidelity, that is, it determines whether viewers can distinguish images from real objects. Fig. 2 shows the result of a subjective assessment of the angular resolution and sense of realness [2]. The higher the angular resolution, the greater the sense of realness, and the sense gently saturates above about 60 cpd. In the experiment, a paired-comparison method was used, and images at six different angular resolutions were presented along with real objects. Participants chose the image that they perceived as better resembling the real object. The experiment setup was such that the effect of factors (e.g. binocular disparity, image size, perspective, luminance, and colour) other than the resolution on the result was minimal.

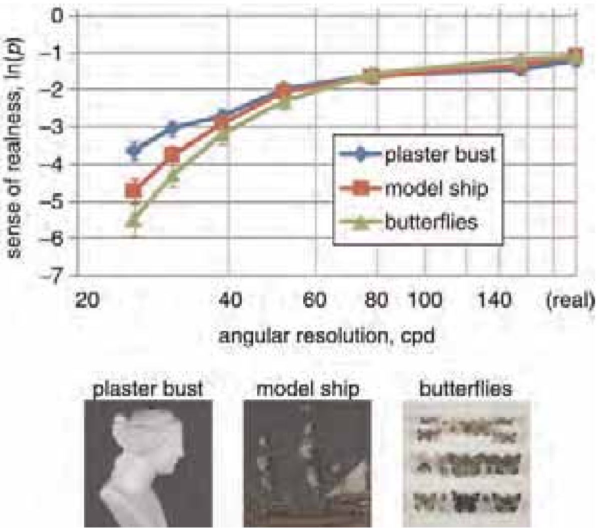

Figure 2 *Angular resolution and sense of realness (mean ± standard error)*

2.3 Spatial sampling parameters of super hi-vision

The spatial resolution of super hi-vision has been determined to be 7840 × 4320, which is four times the resolution of HDTV both horizontally and vertically. Fig. 3 compares three video systems with different spatial resolutions, a 2K system (HDTV), a 4K system, and an 8K system (Super Hi-Vision), in terms of the sense of being there and the sense of realness for a range of FOV angles or viewing distances. The sense of being there is influenced by the FOV; it declines for low-resolution systems at wide FOVs. The sense of realness differs among the three video systems. In the Figure, the angular resolution has been transformed into FOV or viewing distance for the different spatial resolutions. The viewing distance D (H) and the FOV angle θ (deg.) are written as

$$D = 1/V\tan(1/2R)$$

$$\theta = 2\tan^{-1}(8/9D)$$

© The Institution of Engineering and Technology 2011

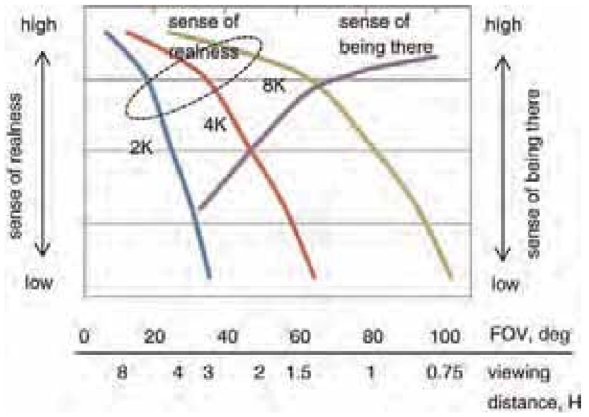

Figure 3 *Comparison of video systems in terms of the sense of being there and sense of realness*

where V is the number of vertical pixels and R (cpd) is the angular resolution at the centre of the screen with the aspect ratio of 16:9.

Super hi-vision can provide both the sense of being there and the sense of realness for a wide range of FOV angles or viewing distances. This feature of super hi-vision is expected to be effectively used in various viewing environments and for large, medium, and small displays. On the other hand, the 4K and 2K systems are effective under certain viewing conditions.

3 Temporal resolution

Motion portrayal is characterised by the perception of motion blur, stroboscopic effect, and flicker. These factors are influenced by temporal video parameters, including the time aperture and frame frequency. The object speeds in programmes also influence motion portrayal.

3.1 Motion blur

Motion blur is caused by the accumulation of incident light in a capture device and/or a hold-type display device associated with eye motion tracking. The time aperture or integration time determines the spatial frequency response (dynamic response), which decreases at high motion speeds. A short time aperture is required for both cameras and displays to improve the dynamic response.

Several experiments have been performed to understand the relationship between motion blur and time aperture. Fig. 4 shows the relationship between time aperture and object speed [3]. The Figure is based on an experiment in which the quality of still images and moving images was assesses for different time aperture–object speed combinations. Only combinations that gave an acceptable degree of motion blur are shown in the Figure. If we assume an object speed of 30°/s, which is typical in HDTV programmes, the time aperture should be in the range 1/200 to 1/300 s.

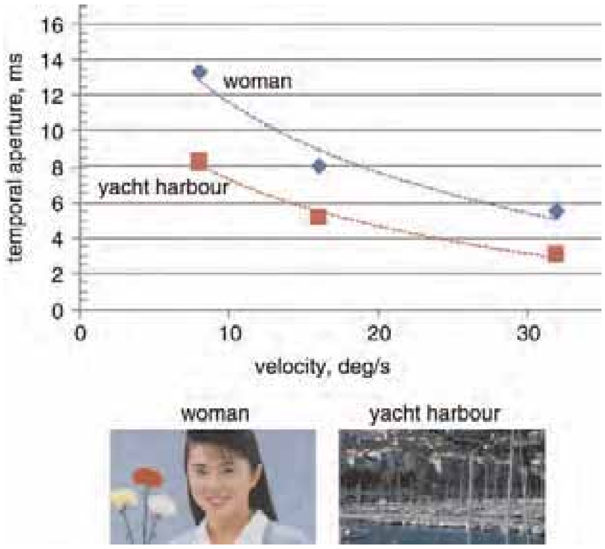

Figure 4 *Motion velocity–temporal aperture combinations corresponding to acceptable degrees of motion blur*

3.2 Stroboscopic effect

The time aperture can be shortened by increasing the frame frequency. The shutter in a camera or black insertion at a display can also be used to shorten the time aperture without changing the frame frequency. However, these techniques may result in the degradation of the picture quality (called the stroboscopic effect or jerkiness), leading to motion being seen as a series of snapshots.

Fig. 5 shows the subjective picture quality in the presence of the stroboscopic effect for different frame frequencies for a time aperture of 1/240 s, which gives an acceptable motion blur [4]. This result suggests that a frame frequency greater than 100 Hz is required for acceptable quality.

3.3 Flicker

Flicker is one type of annoying degradation in moving pictures. A wide FOV in a large screen presentation increases the perception of flicker since human eyes are more sensitive to flicker at the peripheral vision. A short hold time on a hold-type display may also increase the perception of flicker. Fig. 6 shows critical fusion frequencies (CFFs) for two different FOV angles for a 30% time aperture [5]. The Figure confirms that a frame frequency greater than 80 Hz is required for a wide FOV system.

3.4 Temporal sampling parameters of super hi-vision

The above results suggest that the frame frequency of super hi-vision should be at least 120 Hz to achieve a worthwhile improvement in motion portrayal. Naturally, a higher frame frequency would provide better quality, but the improvement tends to saturate.

© The Institution of Engineering and Technology 2011

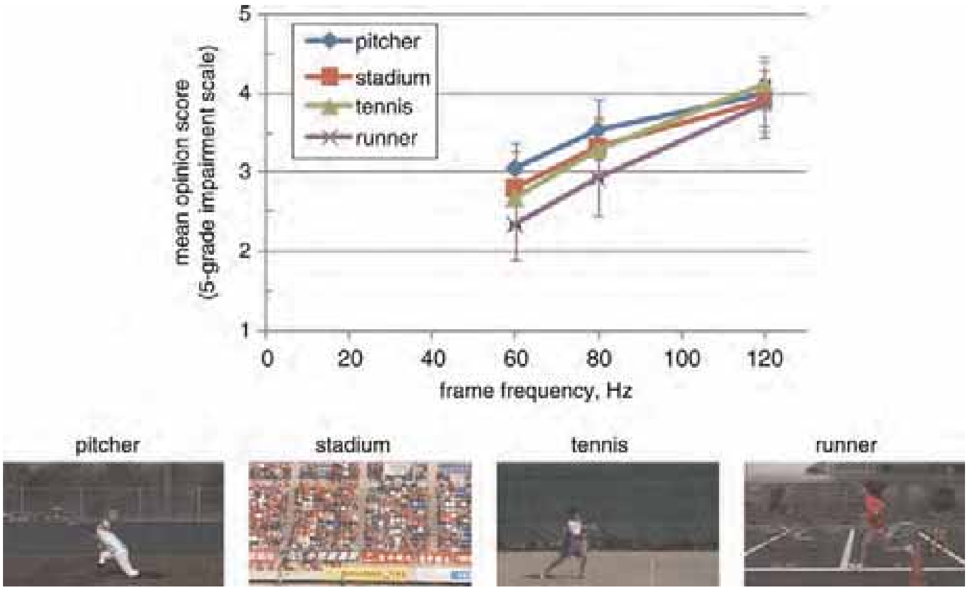

Figure 5 *Picture quality in the presence of the stroboscopic effect at different frame frequencies (mean ± standard error)*

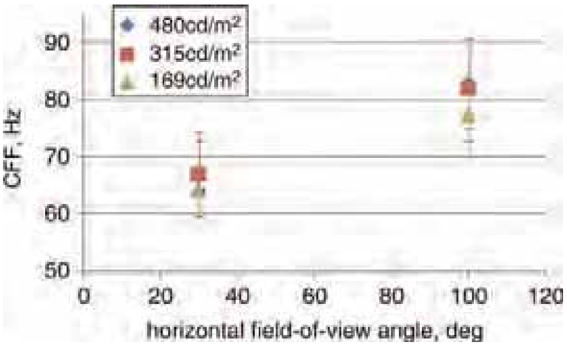

Figure 6 *Plot of CFF against horizontal field-of-view angle (mean ± standard deviation)*

4 Tone reproduction

Discontinuities in tone reproduction, which usually occur as contouring artefacts, should be avoided. This means that quantisation characteristics, particularly the bit depth, should be set such that a modulation corresponding to a one-code value difference between adjacent image areas should be invisible. Fig. 7 shows the contrast sensitivity in a dim surround environment and the modulation transfer characteristics of a gamma 1/2.4 transfer function for 10, 11, and 12 bits. The contrast sensitivity is based on Barten's equation [6], which has been used to determine the bit depth of the D-Cinema distribution master [7]. It is observed that the 11 bit and 12 bit encoding modulation lines are below the visual modulation threshold for the entire luminance range and do not show contouring.

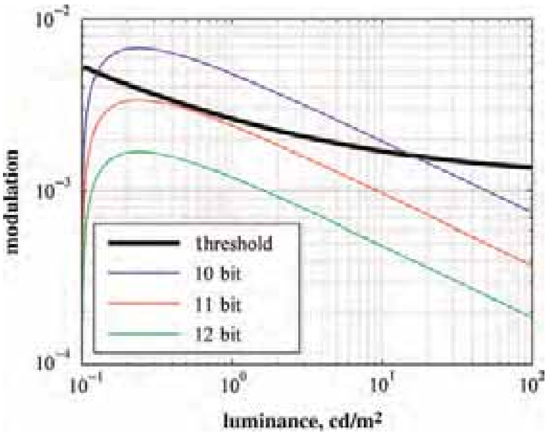

Figure 7 *Modulation threshold and minimum modulation for different bit depths*

5 Colourimetry

Real objects can show highly saturated colours beyond the colour gamut of HDTV. Flat panel displays are becoming capable of reproducing a wider range of colours, and some non-broadcast video systems already handle a wider colour gamut. Super hi-vision should thus cover a sufficiently wide colour gamut, and an efficient and practical method should be devised for this. Requirements for developing a colour representation method and determining parameter values have been defined in terms of target colour, colour coding efficiency, programme quality management, and feasibility of displays.

After comparing several methods for widening the colour gamut while taking the requirements into account, a system colourimetry with RGB monochromatic primaries on the spectrum locus, which can be realised, for example, by using laser light sources in the foreseeable future, has been chosen for super hi-vision [8]. The reference white of D65 remains unchanged. As shown in Fig. 8 and Table 1, the wide gamut colourimetry covers the gamut of HDTV, the

© The Institution of Engineering and Technology 2011

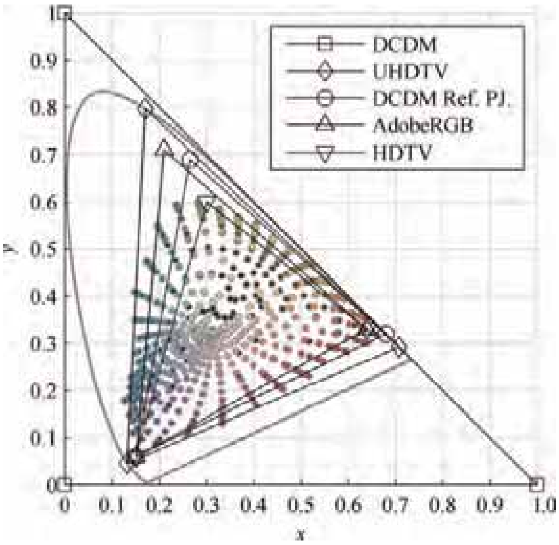

Figure 8 *Pointer's colours and primaries of different video systems*

Table 1 Coverage of Pointer's colours and optimum colour

	Pointer's gamut, %	Optimal colour, %
HDTV	74.4	35.9
Adobe RGB	90.3	52.1
D-Cinema	91.4	53.6
Super hi-vision	99.9	75.8

D-Cinema reference projector, and Adobe RGB, as well as more than 99.9% of Pointer's gamut. Experiments on the capture and display of wide colour gamut images have confirmed the validity of the wide-gamut colourimetry; the texture of objects and highly saturated colours that are closer to those of real objects are reproduced well [9].

5.1 Constant luminance

The representation of video signals in terms of luminance and bandwidth-limited colour-difference signals is an efficient representation method. However, it is known that some luminance information is lost when the luminance signal is derived from gamma pre-corrected RGB signals and the associated colour-difference signals are bandwidth limited. This is called non-constant luminance transmission. HDTV signals using Y′, C'_B, and C'_R suffer from non-constant luminance transmission. We thus propose that the ideal method, i.e. constant luminance transmission in which the luminance signals are derived from 'linear' RGB signals, should be adopted for super hi-vision.

6 Full-spec super hi-vision

On the basis of the intensive studies described above, we have determined the full-spec video parameter values that are suitable for super hi-vision; the values are presented in Table 2. Standardisation is ongoing in ITU-R towards a new Recommendation on Ultra High Definition Television. We hope that the Recommendation will be issued shortly. We are developing a full-spec super hi-vision system based on the specifications.

7 Conclusion

Video parameter values of super hi-vision have been set with the aim of providing an enhanced, or even new, viewing experience to viewers in various environments. Some parameters contribute to an increased sense of being there

Table 2 Basic video parameter values of full-spec super hi-vision

Parameters	Values		
Spatial resolution	7680 (H) × 4320 (V)		
Frame frequency	120 Hz		
Optoelectronic transfer characteristics	$E' = \begin{cases} 4,5E, & 0 \leq E < \beta \\ \alpha E^{0.45} - (\alpha - 1), & \beta \leq E \leq 1 \end{cases}$ $\alpha = 1.0993,\ \beta = 0.0181$		
Bit depth	12 bit		
Primaries and reference white Chromaticity coordinates (CIE, 1931)	R	0.708	0.292
	G	0.170	0.797
	B	0.131	0.046
	D65	0.3127	0.3290
Luminance signal	$Y' = (0.2627R + 0.6780G + 0.0593B)'$		

© The Institution of Engineering and Technology 2011

and to the sense of realness, while others help improve the picture quality by eliminating artefacts in motion portrayal and tone reproduction. Feasibility is also an important factor in determining the parameter values.

Super hi-vision has been demonstrated in mainly theatrical environments. This is because large screen presentation attracts a large audience by offering a strong sense of being there. Since we aim to broadcast services with super hi-vision to homes, 70–100 inch displays viewed at a typical viewing distance of around 2 m, as well as hand-held displays are also considered. The audience can enjoy programmes in a variety of viewing styles; they may opt for a close-up view of an object or to move closer to the display. They would feel that they are watching real objects.

8 References

[1] MASAOKA K., EMOTO M., SUAGWARA M., NOJIRI Y.: 'Contrast effect in evaluating the sense of presence for wide displays', *J. SID*, 2006, **14**, pp. 785–791

[2] MASAOKA K., NISHIDA Y., SUGAWARA M., NAKASU E.: 'Comparing visual realness between high resolution images and real objects'. ITE Technical report, HI2011-62, 2010, pp. 133–135

[3] OMURA K., SUGAWARA M., NOJIRI Y.: 'Evaluation of motion blur by comparison with still picture'. IEICE General Convention, DS-3-3, Japan, 2008, pp. S-5–S-6

[4] OMURA K., SUGAWARA M., NOJIRI Y.: 'Subjective evaluation of motion jerkiness for various frame rates and aperture ratios'. IEICE Technical report, IE2008-205, 2009, pp. 7–11

[5] EMOTO M., SUGAWARA M.: 'Flicker perception for wide-field-of-view and hold-type image presentations'. IDW09, VHF6-3L, Miyazaki, Japan, 2009, pp. 1233–1234

[6] BARTEN P.G.J.: 'Contrast sensitivity of the human eye and its effects on Image quality' (SPIE Optical Engineering Press, Washington, USA, 1999)

[7] COWAN M., KENNEL G., MAIER T., WALKER B.: 'Contrast sensitivity experiment to determine the bit depth for digital cinema', *SMPTE Motion Imaging J.*, 2004, **113**, pp. 281–292

[8] MASAOKA K., NISHIDA Y., SUGAWARA M., NAKASU E.: 'Design of primaries for a wide-gamut television colorimetry', *IEEE Trans. Broadcast.*, 2010, **56**, pp. 452–457

[9] MASAOKA K., OMURA K., NISHIDA Y., *ET AL.*: 'Demonstration of a wide-gamut system colorimetry for UHDTV'. ITE Annual Convention 2010, 6-2, 2010

© The Institution of Engineering and Technology 2011

Selected content from IBC2011

ISSN 2041-5923

ChameleoAD: real-time targeted interactive advertisement for digital signage

C. Jung   R. Tausch   J. Thekkeveettil   T. Oundouh   B. Han
R. Schmidt   D. Wilbers
Fraunhofer-Institut für Graphische Datenverarbeitung IGD, Fraunhoferstraße 5, D-64283 Darmstadt, Germany
E-mail: christoph.jung@igd.fraunhofer.de

Abstract: In this paper the authors present an integrated approach to interactive, real-time targeted multimedia advertisement for digital signage, based on visual audience analysis. Digital signage and digital-out-of-home systems nowadays still employ static content scheduling, and only few of them offer basic interaction capabilities, usually based on touch screens. In this approach the authors use visual sensors attached to a display to measure anonymous properties of the present audience in real-time and to enable touch-less interaction. A combined visual classification scheme analyses facial and full-body image data and integrates results in a long-term observation of individuals. Based on a measured audience profile, considering gender and age, an adaptive media player queries targeted rich media content from a database and presents it to the audience. Content selection is based on an audience state model which considers the spatio-temporal characteristics of the audience (distance, dwell-time) and a content model representing the relations of content and audience. The system automatically offers free-hand gesture interaction to the user, once a certain level of attention is reached. To the authors' knowledge the presented system is the first approach to integrating long-term observation based on facial/full-body visual data, content adaptation and gesture interaction.

1 Introduction

Digital signage has gained more and more importance throughout the last years, especially for the advertisement industry. Sometimes referred to as the '5th screen' in content delivery, public display networks provide efficient means to distribution and presentation of high-fidelity rich-media content on digital displays that are deployed in public areas or public transportation. Usually content playback is scheduled by static playlists, based on heuristics and coarse customer statistics. Some solutions also offer means of interaction, but usually the current audience in front of the display is not considered in content selection. As a consequence presentation of content is often not appropriate for the given audience. Further, non-interactive signage systems can hardly be evaluated after deployment, due to missing customer feedback.

The idea of selecting and composing content in real-time, considering the needs of the present audience, is of course appealing. However, development of adaptive digital signage systems is challenging. If being based on visual sensors, it involves computer vision algorithms, which have to robustly handle various dimensions of the observed scenario (lighting, clutter, crowd, scene dynamics, etc.). Selected content has to be appropriate for the observed audience, reaching it at right time and place. Further, the audience should at some point be addressed directly, to offer interaction or a dialogue, but without being obtrusive.

In this paper we present an integrated approach to real-time targeted interactive advertisement for digital signage, relying on vision-based analysis of the audience, automatic selection of relevant multimedia content and interaction based on free-hands gesture control. To our knowledge we are the first to report on an integrated system that combines advanced visual audience analysis (classifying gender and age on facial and full-body image regions), real-time adaptation of content and interaction based on gestures.

© The Institution of Engineering and Technology 2011

2 Related work

Owing to the integrative nature of our system we summarise related work in the three most relevant areas.

2.1 Adaptive/interactive public displays

Through the last years various authors have presented works on the design of interactive (public) displays. In [1] the authors comprehensively describe basic requirements for designing interactive public displays and specify the design space. A model for the different phases of audience-display interaction (from passing by to interaction and beyond) is presented, similar to the approach in [2]. This model allows specifying the basic behaviour of customers and allows defining appropriate means of adaptation/interaction. The authors identify major challenges of design to (a) attract passers-by, to (b) motivate them to interact and to (c) appropriately deal with social implications of interaction in public. In our work we will especially deal with the first two aspects. Various other works describe approaches to mental models and supported interaction modalities. The ReflectiveSigns system in [3] uses cameras and face detection to measure view time of the audience. The system learns preferred content by measuring dwell times, to improve selection for subsequent audience. It is reported that view time is almost not affected by selected (optimised) content, but from location. Unlike in our system no characteristics like gender or age are used for content selection. CityWall [4] offers collaborative multi-touch interaction to customers in a field trial installation. The study points out that observing users in front of a display can attract additional users. A usual problem in interactive display applications is how to make customers aware of the interactive capabilities. It is demonstrated that explorative social learning can be more effective than manuals or tutorials. In a rather early work the authors of [2] present an approach based on gesture recognition, requiring wearable reflective landmarks for visual body tracking, but which would of course not be suitable for our public scenario. In [5] a digital column based on projection is presented. The authors claim that user tracking is suitable for implicit interaction and to get users involved. However, face detection alone is not sufficient, as users are not facing the display continuously.

2.2 Visual classification of gender/age

Visual classification of humans can be performed using cameras or comparable visual sensors. Most researchers classify human visual traits based on facial or full-body image regions (showing a person in upright pose, e.g. in front of a signage display), which have been previously detected/tracked using computer vision algorithms. Analysis of facial regions usually results in better classification rates than full-body, but obviously facial regions are not always available. Therefore both information sources have to be considered (and combined), if the audience will be continuously analysed while being present at the display. The authors in [6, 7] perform age classification on human facial regions, usually based on frontal faces, but requiring prior automatic or manual alignment of face regions. Age classification is usually able to distinguish certain age groups (e.g. young, adult, elderly) or delivers continuous values (as for some facial approaches). Golomb *et al.* [8] for the first time performed gender recognition on face images. It was followed by many other approaches based on computer vision and machine learning techniques [9, 10], resulting in high recognition rates of 86% and more. There have been mainly two publications so far studying full-body recognition of visual traits (gender only) [11, 12]. In our previous work [13] we further could show that a continuous observation of people, by temporal integration of per-frame classification results, can remarkably improve overall classification of tracked individuals.

2.3 Gesture-based interaction

Gesture recognition from visual data has been studied since many years in computer vision, both for touch-based as well as free-hands interfaces. In [14] a good overview of known approaches is provided. Recently, Microsoft hit the mass markets with its *KINECT* sensor and APIs, enabling tracking a person's body and recognising gestures, which turned out to be a massive boost for the human machine interfacing research community. Besides using standard digital cameras (monocular, stereo or multi-view), also structured light and time-of-flight (ToF) cameras have gained much importance through the last 6 years. Both have in common that they deliver range (depth) images, which provide the distance of objects to the camera for each pixel. [15] provides detailed background knowledge on the technology and related applications. The authors in [16] use a ToF camera to recognise static finger configurations (hand postures). An approach to recognition of deictic gestures (pointing) is presented in [17]. This approach is similar to our technique, but it requires head position to determine the screen position as the extended line between head and hand.

3 Overview and system architecture

Our proposed system enables (a) real-time adaptation of advertisement content to the present audience and (b) collecting audience and content information in a database to perform audience measurement. The basic architecture (see Fig. 1) comprises sensing components for visual audience analysis and gesture interaction, a web-browser-based adaptive player, database services for collecting audience profiles and a tool to analyse and visualise collected data for the operator. The sensing component uses visual sensors and computer vision algorithms to acquire the scene in front of the display. Persons are automatically detected, tracked and classified (anonymously) into gender, age and status (orientation, distance, attention). The recognised audience

© The Institution of Engineering and Technology 2011

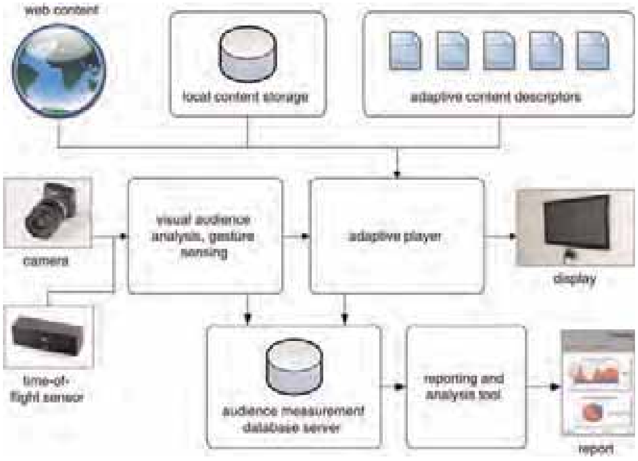
web content
local content storage
adaptive content descriptors
camera
visual audience analysis, gesture sensing
adaptive player
display
time-of-flight sensor
audience measurement database server
reporting and analysis tool
report

Figure 1 *ChameleoAD basic system architecture*

profile is sent to the adaptive player that controls the digital signage display. Based on the profile, the player updates an internal state machine and queries a content description for relevant content. The description scheme is based on XML and encodes the relations among content items (e.g. a video plus a website as part of a cross media clip) and between content and audience (e.g. a video being tagged as 'relevant for female'). An HTML5 browser component renders the content at the given URL and manages transitions among subsequent content items. While the system is presenting adaptive content, anonymous data about the audience is stored in a database, to generate customised reports about customer and usage statistics.

4 Vision-based audience analysis

Vision-based analysis in our system uses cameras and visual sensors to continuously observe people in front of a public display, in order to automatically recognise anonymous properties like gender, age or level of attention. The extracted audience properties are then transmitted to the adaptive player component for selection of appropriate content. The system is based on previous work we presented in [13].

Analysis based on visual data is highly challenging, mainly owing to large variations of individuals, in terms of body pose, physique, scene geometry, illumination, clothing style and inter-object occlusion. Our vision-based analysis subsystem receives a video image stream from a standard uEye UI-1485LE-C digital camera attached to the digital signage display (see Fig. 2). A person detector component analyses the images and determines the location of pedestrians (see Fig. 3 left). The detector is based on HoG image features, describing appearance of an object based on image gradient distributions, and support vector machines (SVM), a well-known classification scheme in machine learning. Based on the detections and additional foreground segmentation, a tracking based on Kalman filtering is initialised for each person detected. The tracking ensures that the person can be followed throughout a sequence of images, even if observation based on detection fails in particular images. Further the system applies face detection and tracking on the images, to locate facial regions of persons (see Fig. 3 right).

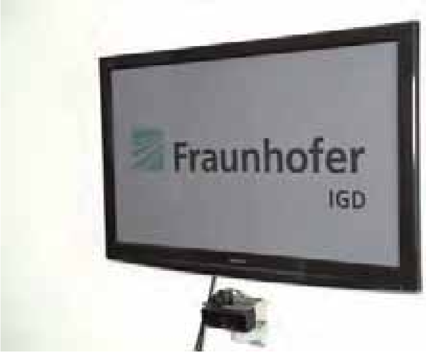

Figure 2 *Display with uEye camera and Panasonic ToF range sensor mounted below*

Based on the detected facial and full-body regions of persons a classification scheme determines characteristics of that individual. Currently recognition of gender (male, female) and age group (young, middle, elderly for full-body; 10–20, 20–30, 40–50, 50+ for facial) is realised. The classification is also based on SVM, and employs HoG image features for full-body and the facial regions.

The independent classification results for each facial and full-body region in each image are further integrated over time, to result in a combined classification result for a tracked individual. This approach stabilises the classifications for a whole track, as it filters false classifications. In [13] we could show that this track-based observation can improve classification performance by up to 10% per tracked person. Currently classification achieves up to 80% for full-body and 90% average precision for facial frontal images (gender).

5 Free-hands gesture interaction

When a person has been attending an advertisement on the display for a while, being near to the display, the system offers explicit interaction. As the display should be accessible without touching (e.g. if placed in a shopping window, or to avoid screen contamination), a free-hands interaction approach was chosen. In the interaction state a person can use hands to move a cursor on screen and click, similar to a computer mouse.

The free-hands interaction component uses range (depth) data from a Panasonic EKL 3104 ToF camera mounted below the display (see Fig. 2). The ToF camera illuminates the audience with invisible IR light pulses and measures the time a pulse requires to propagate from the camera to the object, to compute the distance. For more details on

© The Institution of Engineering and Technology 2011

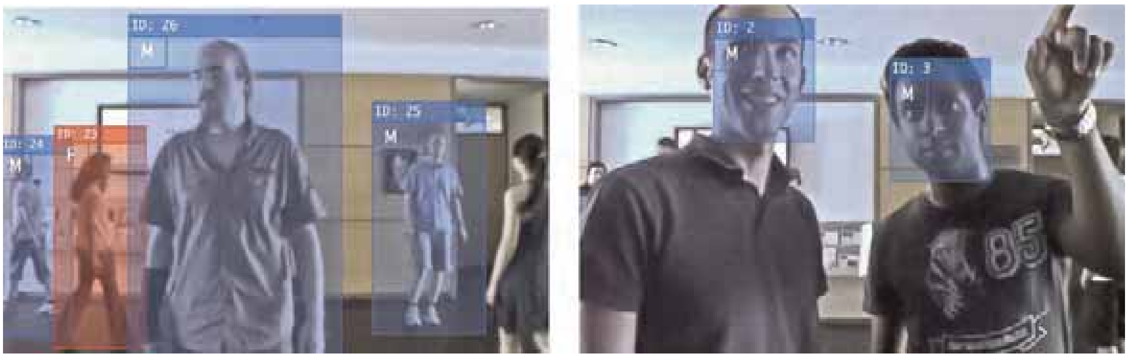

Figure 3 *Gender classification on full-body (left) and facial (right) image regions*

ToF and applications please refer to [15]. Our ToF sensor delivers a range image and an IR intensity image (see Fig. 4). Besides ToF, the system has also been tested with Microsoft's *KINECT* camera.

The interaction component at first captures depth images as well as the intensity images frame-wise from the ToF camera. After removing noise, clutter and unnecessary background, the hand location is extracted based on depth information from the user's interaction area (see Fig. 4 right). This area is dynamically adapted relative to the user's upper body geometry and head/face position, which is localised in the intensity image. Our approach is related to that in [17], but it does not compute the cursor location on screen as the extended line from head to hand, but just as the hand position relative to the person's body. Currently, our system supports recognition of wiping hand gestures that later will be used e.g. for navigating in list content.

6 Adaptive and interactive media player

The adaptive player receives data from the audience analysis (AA) and gesture interaction component. Based on an audience state model, a player state model and a content model, the player can query suitable content that considers the current state of the audience. In the following we will describe the models and their interplay.

6.1 Audience model

Usually, digital signage displays are deployed at locations with specific audience characteristics [18], like point-of-transit (PoT), point-of-wait (PoW) or point-of-sales (PoS). In our approach we want to focus on PoT and PoW locations. In these cases people either pass by the display (e.g. pedestrian precinct), or stand somewhere waiting (e.g. bus station). To reflect this characteristic in our model (see Fig. 5), it was chosen to define a state (1) where a person is at a certain distance to the display, either passing by or waiting. In this state the person is only visible as full-body person in the camera, and first gender/age classification on this data is performed. It should be noted that this aspect differentiates our approach from most existing solutions, which only rely on facial data. When the person is approaching the display,

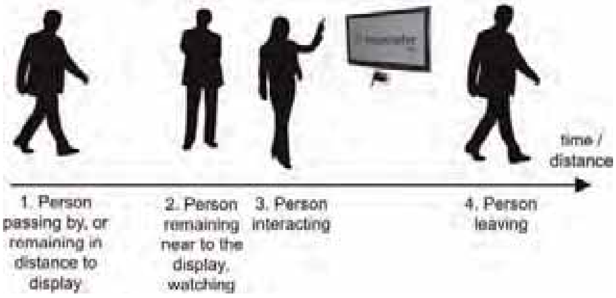
time / distance
1. Person passing by, or remaining in distance to display
2. Person remaining near to the display, watching
3. Person interacting
4. Person leaving

Figure 5 *Audience state model*

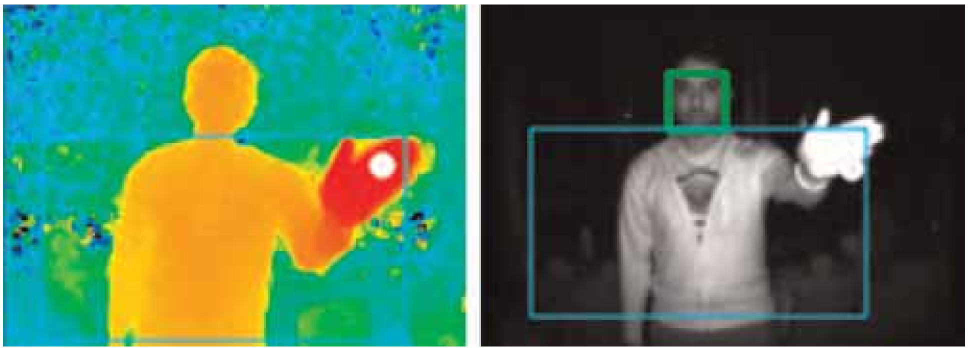

Figure 4 *Free-hands interaction based on face-/hand tracking and gesture recognition (left: range image showing tracked hand, right: face detection and cursor area; image resolution is limited by ToF sensor)*

© The Institution of Engineering and Technology 2011

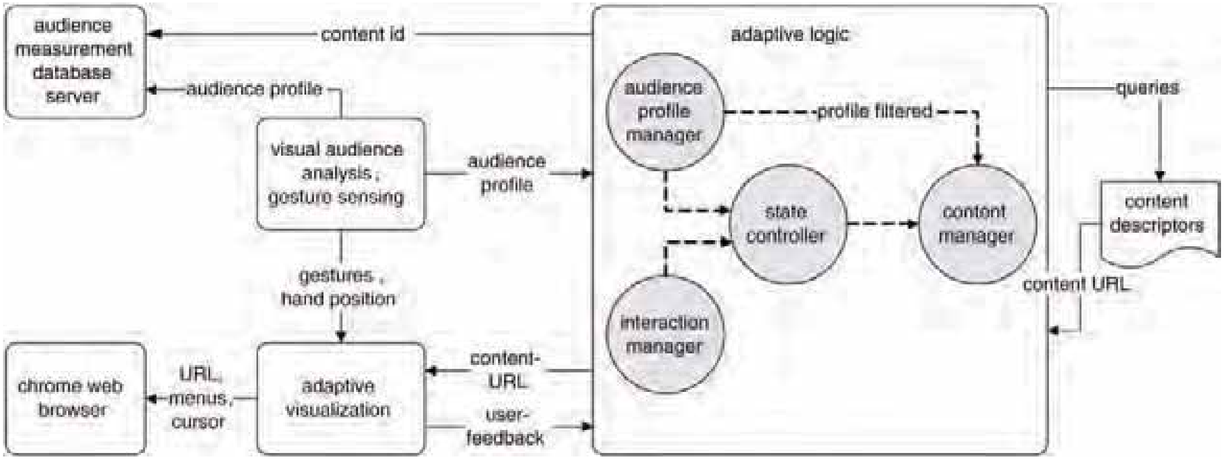

Figure 6 *Adaptive Logic reacts to audience profile changes, selects content and controls player web browser front end*

and remains at closer distance, either watching or waiting (state 2), the system starts combining full-body classification with facial data. At that time, the observation time (dwell time) of each detected and tracked individual is recorded, to have a simple indication for level of attention. After a certain threshold of attention time has been reached, the player presents a menu to the user, offering interaction on the display (state 3). The user can now interact with the presented content, explicitly quit the interaction session or simply leave the area in front of the display (state 4).

6.2 Player model

The adaptive player in ChameleoAD is composed of three major components (see Fig. 6), the adaptive logic (AL), adaptive visualisation (AV) and an instance of the Google chrome web browser (WB). AL is the core element of the player, it receives audience profiles from the audience analysis (AA) component and explicit user feedback from AV. Based on its internal functionality, it generates XML queries that are applied on the XML content descriptors. AV basically acts as a browser controller module. It gets new content (as HTTP URLs) from AL and loads them in WB, or shows interactive menus and draws the user's cursor.

The AL component is the main controller in the adaptive player. Its audience profile manager stores incoming profiles and performs filtering. This is necessary to avoid too frequent state changes, which could cause the player to switch content permanently. Depending on the chosen filter length in time, the system usually is able to switch content in about 1 second (depending on network connection). The interaction manager receives user input such as the result of the 'offer interaction' dialogue, which could be refusal or acceptance. Both inputs influence the

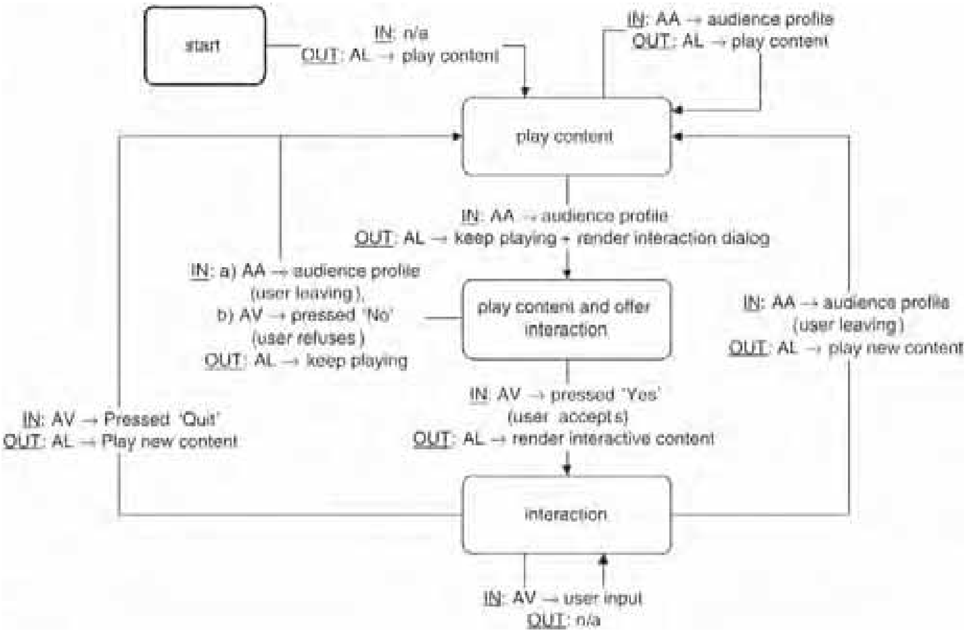
start
IN: n/a
OUT: AL → play content
IN: AA → audience profile
OUT: AL → play content
play content
IN: AA → audience profile
OUT: AL → keep playing + render interaction dialog
IN: a) AA → audience profile
(user leaving),
b) AV → pressed 'No'
(user refuses)
OUT: AL → keep playing
play content and offer interaction
IN: AA → audience profile
(user leaving)
OUT: AL → play new content
IN: AV → pressed 'Yes'
(user accepts)
OUT: AL → render interactive content
IN: AV → Pressed 'Quit'
OUT: AL → Play new content
interaction
IN: AV → user input
OUT: n/a

Figure 7 *Player state machine*

© The Institution of Engineering and Technology 2011

state controller, which maintains transitions in the player's state machine. The content manager generates XML queries from the current audience profile (and currently played content, if required), and retrieves content (URL) from the XML repository.

The core of the AL component is represented by a finite state machine, consisting of three elementary states (see Fig. 7) the player can take during operation, 'play content', 'play content and offer interaction' and 'interaction'. State transitions are triggered either by changes in the audience profile, for example a person is leaving, or showing a certain level of attention. It should be noted here that a change in audience composition does not necessarily call for a state change. The player remains in playback state, but changes type of content which can be considered as a parameter of this state.

6.3 Content model

The basic idea of the chosen content model was to separate individual atomic content items (like videos, webpages, etc.) from the composition or order they will be played during adaptation. The selection process for targeted content is based on a query on a previously created XML-based description. In this description, all available content items are listed and annotated with additional meta data (Table 1), e.g. related gender, age groups, order, content group, audience distance, etc. The result of a query can therefore be any type of content, related to the state of the present audience and the previous content shown. The system can handle various types of multimedia content that can be rendered in recent web browsers, ranging from images, HD video content, web pages to complex browser games and interactive flash content. Content is stored as a URL in the XML representation, and can be located either locally or on a web server.

A typical sequence of content selections could be the following: a person enters the sensing field of the system; the player selects a video (type 'far distance') for playback, suitable for the person's gender and age group; the person is attracted by the video, which mainly focuses on highlight scenes, with large-font on-screen messages, and approaches the display; the person remains in front of the display for a while, while the player detects the person's face and switches to near-distance content; the selected video contains more detailed on-screen messages, which could now be read by the person; after a while the system offers interaction to the user (still playing content); if the user accepts, interactive related content is shown, like e.g. games or a ticket ordering service.

Table 1 Content annotations relevant for adaptation. The player supports further metadata like genre, title, etc.

Target gender:	male, female
Target distance:	far, near
Target age group:	10–20, 20–30, 30–40, 40–50, 50 + years
Interaction level:	interactive, non-interactive
Content type:	web page, video, image

7 Conclusion

In this paper we presented an integrated approach to adaptive and interactive digital signage. To our knowledge this is the first time that long-term observation by visual sensors, classification based on facial and full-body images, gesture interaction and adaptation have been integrated in one system. We could show how the interplay of detection, tracking, content adaptation, dialog and interaction can be realised, considering basic models of public display audience.

Although there are already studies known in literature which report on digital signage and interaction, our combination of adaptation and interaction must be considered individually. We plan to conduct a first experimental study in a real world setting, to evaluate our integrated approach with a public audience. Further we will improve our audience models, to represent more characteristics of persons in a scene (like clothing style or social relations), which would in turn enable new adaptation rules. Concerning gestures interaction, we will integrate recognition of a larger amount of hand gestures, and also enable head gestures like nodding and shaking. As we are convinced that gestures recognition should only be one possible option for the user, we will also investigate multi-modal interaction, e.g. based on speech recognition.

8 References

[1] MÜLLER J., ALT F., SCHMIDT A., MICHELIS D.: 'Requirements and design space for interactive public displays'. MM'10, Firenze, Italy, 25–29 October 2010

[2] VOGEL D., BALAKRISHNAN R.: 'Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple user'. Proc. UIST '04, 2004, pp. 137–146

[3] MÜLLER J., EXELER J., BUZECK M., KRÜGER A.: 'ReflectiveSigns: digital signs that adapt to audience attention'. Proc. Pervasive 2009, Nara, 2009

[4] PELTONEN P., KURVINEN E., SALOVAARA A., JACUCCI G., ILMONEN T., EVANS J., OULASVIRTA A., SAARIKKO P.: 'It's Mine, Don't Touch!: interactions at a large multi-touch display in a city centre'. CHI '08, 2008, pp. 1285–1294

[5] BEYER G., ALT F., KLOSE S., ISAKOVIC K., SAHAMI SHIRAZI A., SCHMIDT A.: 'Design space for large cylindrical screens'. Proc. Third

© The Institution of Engineering and Technology 2011

Workshop on Pervasive Advertising and Shopping, Helsinki, Finland, 2010

[6] KWON Y.H., VITORIA LOBO N.D.: 'Age classification from facial images', *Comput. Vis. Image Understand.*, 1999, **74**, (1), pp. 1–21

[7] LANITIS A., CHRISTODOULOU C.: : 'Comparing different classifiers for automatic age estimation', *IEEE Trans. Syst., Man, Cybern.*, 2004, **34**, (1), pp. 621–628

[8] GOLOMB B.A., LAWRENCE D.T., SEJNOWSKI T.J.: 'Sexnet: A neural network identifies sex from human faces'. Proc. NIPS-3 (1990), pp. 572–577

[9] BUCHALA S., DAVEY N., GALE T.M., FRANK R.J.: 'Principal component analysis of gender, ethnicity, age, and identity of face images'. IEEE ICMI, 2005, vol. 2

[10] MAKINEN E., RAISAMO R.: 'Evaluation of gender classification methods with automatically detected and aligned faces', *IEEE TPAMI*, 2008, **30**, (3), pp. 541–547

[11] CAO L., DIKMEN M., FU Y., HUANG T.S.: 'Gender recognition from body'. Proc. MM '08, New York, NY, USA, 2008, pp. 725–728

[12] COLLINS M., ZHANG J., MILLER P., WANG H.: 'Full body image feature representations for gender profiling'. VS '09 ICCV: IEEE Workshop on Visual Surveillance, 2009, pp. 1235–1242

[13] JUNG C., TAUSCH R., WOJEK C.: 'Real-time full-body visual traits recognition from image sequences'. 15th Int. Workshop on Vision Modeling and Visualization (VMV) 2010, Siegen, November 2010

[14] MITRA S., ACHARYA T.: 'Gesture recognition: A survey', *IEEE Trans. Syst., Man Cybern., Part C (Applications and Reviews)*, 2007, **37**, (3), pp. 311–324

[15] KOLB A., BARTH E., KOCH R., LARSEN R.: 'Time-of-flight sensors in computer graphics'. Eurographics 2009 – State of the Art Reports, March 2009, pp. 119–134

[16] KOLLORZ E., HORNEGGER J.: 'Gesture recognition with a time-of-flight camera'. DAGM '07, 2007

[17] HAKER M., BÖHME M., MARTINETZ T., BARTH E.: 'Deictic gestures with a time-of-flight camera'. International Gesture Workshop GW, 2009, (*LNAI*, **5934**), pp. 110–121

[18] KELSEN K.: 'Unleashing the power of digital signage – content strategies for the 5th screen' (Focal Press, 2010)

© The Institution of Engineering and Technology 2011

Selected content from IBC2011

ISSN 2041-5923

Millimetre-wave active imaging system using 60 GHz band

H. Kamoda J. Tsumochi F. Suginoshita

Science & Technology Research Laboratories, Japan Broadcasting Corporation (NHK), 1-10-11 Kinuta Setagaya-ku, Tokyo 157-8510, Japan
E-mail: kamoda.h-ci@nhk.or.jp

Abstract: Millimetre waves can penetrate optically opaque substances, such as smoke and wood, so using millimetre waves instead of visible light will enable objects obscured by these substances to be viewed. The prototype of a millimetre-wave active imaging system for use in broadcasting is described. The prototype transmits 60 GHz millimetre waves towards a target scene and receives reflected waves from objects in the scene through a phased array antenna that enables the antenna beam to be rapidly scanned. This active prototype can analyse the time-of-flight of the received waves to obtain the range information of the scene; it can produce a 3D profile of the scene by performing time-of-flight analysis for each position of the antenna beam. In addition, detection of a slight movement along the range dimension is possible by observing the phase change of the reflected waves.

1 Introduction

The use of millimetre waves (radio waves ranging from 30 to 300 GHz) enables imaging through different opaque media, such as smoke, clothing, and certain building materials [1]. Recent advances in millimetre-wave technologies have been leading the development of millimetre-wave imaging systems for many applications. In particular, significant attention is being drawn to millimetre-wave body scanners for security screening of people at airports to detect weapons hidden under clothing in response to the increased threat of terrorism [1].

There are two types of millimetre-wave body scanners: one is passive, capturing and visualising millimetre waves that are naturally emitted from the human body and objects [2], and the other is active, illuminating the human body with millimetre waves and receiving and visualising the reflected waves [3]. These scanners are used in close proximity to the target human body and in a fixed environment, so the imaging performance, such as image resolution, is optimised to detect concealed weapons.

We are exploring the possibility of extending millimetre-wave imaging technology to broadcasting applications. As mentioned above, millimetre-wave imaging enables visualisation of humans, objects, or anything that reflects millimetre waves, that are obscured by not only clothing but also smoke, wood, plastic, etc. which is not possible with ordinary optical cameras. There are many possible scenarios for use of millimetre-wave imaging technology, e.g. emergency reporting from a site where a building on fire is obscured by smoke or flames, or a hostage situation within a house where curtained windows prevent a view inside. Other scenarios include use for nature programmes, where an millimetre-wave imager can observe wild animals that are obscured by foliage and so on.

In light of this broad range of possible scenarios, there are several requirements that are different from those for millimetre-wave body scanners for security screening: a) the frame rate must be reasonably high enough, preferably at a video rate; b) variable focusing capability is required as the depth of the scene may be tens of metres; c) features to help users to recognise the obscured scene better than with conventional two-dimensional (2D) intensity-based images are desirable because the image resolution would obviously be very low compared to ordinary optical camera images and the imaging will be done without prior knowledge of what kinds of objects are in the scene.

A prototype millimetre-wave imaging system that meets these basic requirements is described. The imaging test

© The Institution of Engineering and Technology 2011

results obtained with the prototype are also presented and discussed.

2 System design

To meet the requirements described in the previous Section, active imaging architecture was used, for reasons detailed later. A system block diagram of the prototype millimetre-wave imaging system is in Fig. 1. The prototype comprises a transmitter, receiver, antenna, directional coupler, and display. A 60 GHz band millimetre wave was chosen for this prototype. A digital signal processor implemented in a field-programmable gate array is used for the baseband signal processing to enable fast processing. The directional coupler enables the single antenna to be shared by the transmitter and the receiver. The principle of operation is based on a frequency-modulated continuous wave radar [4]. This principle is detailed next.

The transmission signal, which is a chirp signal with sweeping bandwidth of 450 MHz, is generated in the digital signal processor. Then, the baseband chirp signal is up-converted and frequency-multiplied such that the transmitted RF signal (millimetre-wave signal) has a chirp signal sweeping a 1800 MHz bandwidth with a centre frequency of 60.925 GHz. The millimetre-wave signal is then transmitted towards the scene to be imaged through the antenna, which has a very narrow beam. The objects in the scene reflect the millimetre-wave signal back to the antenna. The received millimetre-wave signal enters the receiver through the directional coupler, and the down-converter mixes the received signal with the transmitted signal, producing a beat signal that has the difference frequency (beat frequency) between the transmitted and received signals. The beat frequency depends on the range (distance) of the object that reflected the millimetre-wave signal because the beat frequency arises from the time delay of the received signal, as illustrated in Fig. 2. Therefore, Fourier transform is performed on the beat signal to obtain the profile of the scene along the range dimension, which is called range profile. This series of processes is repeated while the narrow antenna beam is raster scanned two-dimensionally in cross-range dimensions. Thus, the 3D profile of the scene can be obtained and is finally presented on the display. Such 3D profiles of the scene may enable better understanding and recognition of the scene than conventional simple 2D images with no range information.

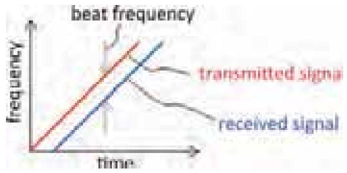

Figure 2 *Beat frequency arising from time delay of received signal*

The antenna beam must be rapidly scanned to achieve a high frame rate. Therefore, an electronic scanning antenna was newly developed at the 60 GHz band. This antenna can also focus its beam at arbitrary ranges.

The details of the electronic scanning antenna, circuit architecture for sharing the single electronic scanning antenna, and the process to obtain and visualise the 3D profiles of scenes are described in the following.

2.1 Electronic scanning antenna

The antenna used in millimetre-wave imaging plays a very important role because it is equivalent to the lens of an ordinary optical camera. In this system, the antenna also has to focus and scan its beam very rapidly. The aperture size determines the beam width, which is related to the angular resolution, i.e. to obtain reasonably high resolution, a large aperture is necessary. Hence, an electronic scanning antenna with a large aperture at the 60 GHz band is required. Achieving this is very challenging and, to the best of the authors' knowledge, such a large electronic scanning antenna has not yet been discussed in the literature. Therefore, we developed a new electronic scanning reflectarray antenna (ESRA).

A schematic of the ESRA is in Fig. 3. Reflectarray antennas are a type of phased array antenna. They consist of small reflecting elements, each equipped with a phase

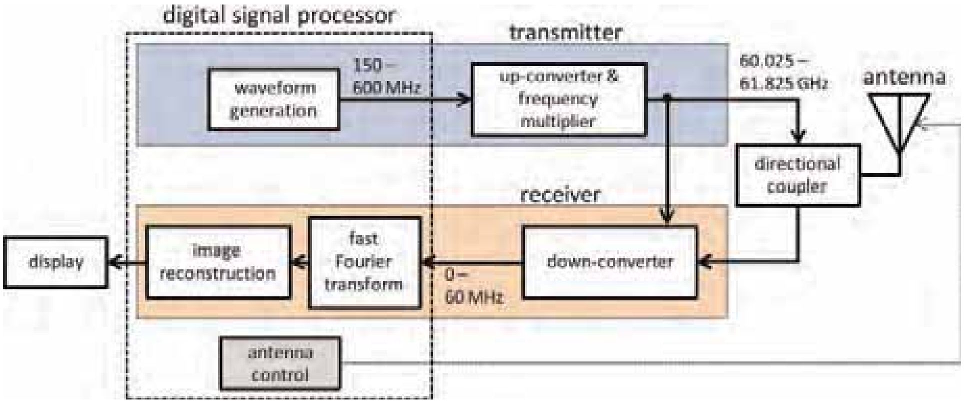

Figure 1 *System block diagram*

© The Institution of Engineering and Technology 2011

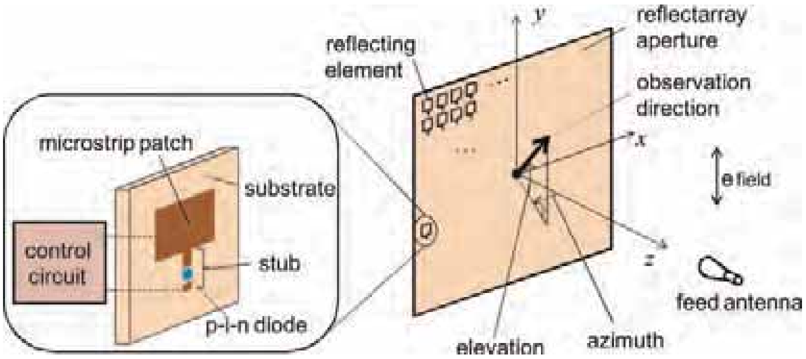

Figure 3 *Electronic scanning reflectarray antenna (ESRA)*

shifter and a feed antenna to spatially feed the reflecting elements. They have no complicated feeding circuits causing ohmic losses [5], as opposed to conventional phased array antennas.

The remaining challenge is how to implement the phase shifters on the considerable number of reflecting elements. The inset of Fig. 3 illustrates a reflecting element with a phase shifter that was used for the ESRA. It consists of a microstrip patch and a stub loaded with a p-i-n diode. The reflecting elements are formed on a printed circuit board (PCB), and the control circuit for the p-i-n diode is placed behind the PCB, which sets the p-i-n diode state to either ON or OFF. The reflection phase of the reflecting element can be changed by the state of the p-i-n diode. That is, a single-bit phase shifter is included in the reflecting element with a very simple structure. This simple structure enabled the construction of the large ESRA.

A photograph of the fabricated ESRA with aperture size of 575 × 575 mm is in Fig. 4. To verify the beam scanning capability, every p-i-n diode across the aperture is independently controlled, such that the desired phase front or beam direction is obtained. The measured beam patterns in the azimuth plane when the beam direction was set to every five degrees in the azimuth plane are shown in Fig. 5. Note that all the beam patterns were superimposed in this Figure. Similar results were obtained for the beam patterns in the elevation plane. The measured antenna gain was 41 dBi. The beam widths were approximately 0.6°; they are a relevant factor in determining the angular resolution of the imaging system. The beam was confirmed to be successfully scanned. Fig. 5 shows the beam patterns with a focus at infinity, but similar patterns were also obtained for the ranges of 1.5, 3, 5, and 10 m when the beam was focused at the respective ranges. Consequently, it was confirmed that scanning and focusing can be performed completely electronically.

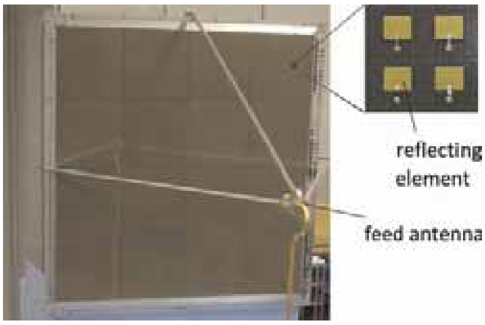

Figure 4 *Fabricated ESRA*

2.2 Sharing electronic scanning antenna between transmitter and receiver

In active imaging, the two-way beam pattern, which is the product of the beam patterns of the transmitter and receiver antenna, determines the image quality of the imaging system. Sharing a single large antenna between the transmitter and receiver produces a two-way beam pattern with an effectively lower side-lobe level and narrower beam width without using separate two equal-sized large antennas.

On the other hand, sharing a single antenna worsens the isolation between the transmitter and receiver because part of the transmitting signal directly leaks into the receiver. In general, the power of the leakage signal is significantly larger than that of the received signal that has undergone free space propagation. That is, the leakage signal desensitises the receiver so that a weak signal can no longer be detected.

To solve the isolation problem, we employed single-ended mixer architecture to exploit the leakage signal, i.e. to use the

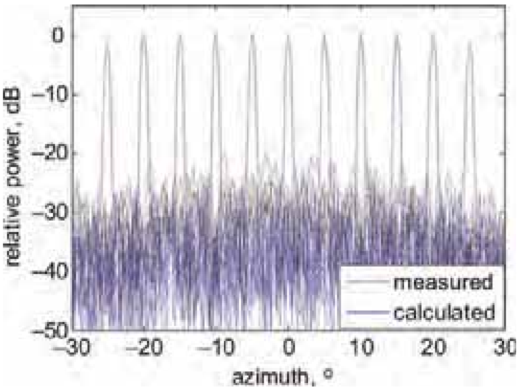

Figure 5 *Beam patterns of ESRA*

© The Institution of Engineering and Technology 2011

leakage signal as a reference signal to obtain the beat signal, which is illustrated in Fig. 6. As the Figure shows, there is actually no need for the transmission line to feed the transmitting signal to the receiver, which was illustrated in Fig. 1. Only a single-ended mixer is needed to produce the beat signal from the mixture of the leakage signal and the received signal. In such a way, a simple circuit taking advantage of the leakage signal was used for the prototype millimetre-wave active imaging system to achieve sharing of the ESRA between the transmitter and receiver.

2.3 Process to obtain and visualise 3D profile of scene

The field-of-view of the prototype was defined as 40° (azimuth) × 30° (elevation) × 16.7 m (range). The sampling interval in the angular dimension was set to 0.43°, and that in the range dimension was set to 8.3 cm.

Subsequently, the beam of the ESRA is raster scanned for every 0.43°. At each antenna beam position, transmission and reception of a millimetre-wave signal is performed. The received signal is then processed to obtain the range profile. The range profile is sampled at every 8.3 cm up to 16.7 m, for a total of 200 points. Thus, the 3D profile of the scene consists of 1.2696 million sampled points, or voxels (volumetric pixels). Note that every voxel holds the amplitude and phase of the received signal.

Then, each voxel is given a brightness value depending on the power of the received signal. A threshold with respect to the power of the received signal is applied to all the voxels to eliminate voxels under a specific noise level. The remaining voxels are displayed on a PC screen using directX technology, so that a perspective view of the 3D profile of the scene is seen from an arbitrary viewpoint in real time.

One advantage of active imaging other than the capability for obtaining a 3D profile of a scene is that the phase data can be used to detect any slight movement of objects along the range dimension by comparing the phase of a particular voxel with that of the same voxel in the previous frame. The wavelength of a 60 GHz wave is 5 mm, so only 2.5 mm movement results in 360° phase change, which makes it possible to detect slight movement of even less than 1 mm. Taking advantage of this property, the prototype also includes a feature to display the movement detection results on the 3D profile. This may also help users to recognise moving objects, which may be living creatures, and to distinguish them from the surrounding still objects.

3 Experimental results

First, the frame rate was measured to be about 2 frames per second. This is not yet video rate, but it is high enough to recognise relatively slow movement of humans and so may be suitable for most cases. Note that the frame rate can be readily increased to up to 4 frames per second at the expense of a lower signal-to-noise ratio of the image. The factor limiting the maximum frame rate is currently the ESRA's speed of scanning.

Next, 2D images in the azimuth elevation plane were taken to verify the variable focusing capability. The focused ranges tested were 1.2, 1.6, 2.3, and 3.5 m. The images were confirmed to be clear for objects at the respective focused ranges. Then, the effect of sharing the ESRA between the transmitter and receiver was also investigated with the 2D images. We had previously developed a prototype with a separate transmitter and receiver antenna, where a small horn antenna (14 dBi) for the transmitter and the ESRA for the receiver were used [6]. Images taken by the previous prototype and the current prototype were compared to determine the effect of sharing the ESRA. Images of a plastic human mannequin taken at the range of 2.3 m are shown in Fig. 7. The left image was taken by an ordinary optical camera. The middle and the right are millimetre-wave images. In the middle image, taken by the previous prototype, background noise is high because of the higher side-lobe level of the two-way beam pattern, while the right image, taken by the current prototype, has lower background noise and shows a clearer outline of the mannequin. In this case, the background noise level in the right image was improved by approximately 10 dB compared with the middle image. Thus, it was confirmed

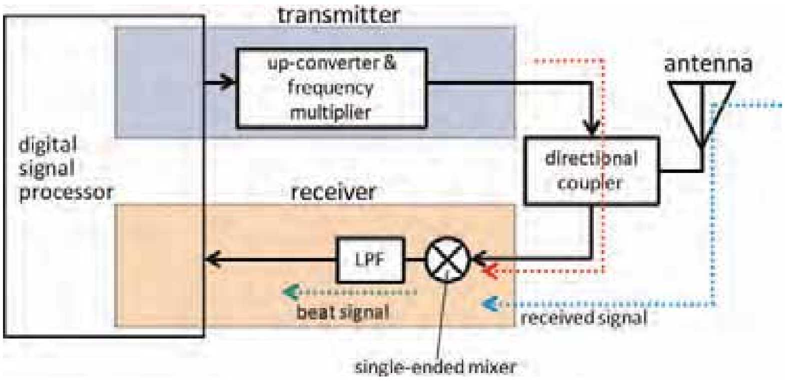

Figure 6 *RF circuit diagram of sharing single antenna between transmitter and receiver*

© The Institution of Engineering and Technology 2011

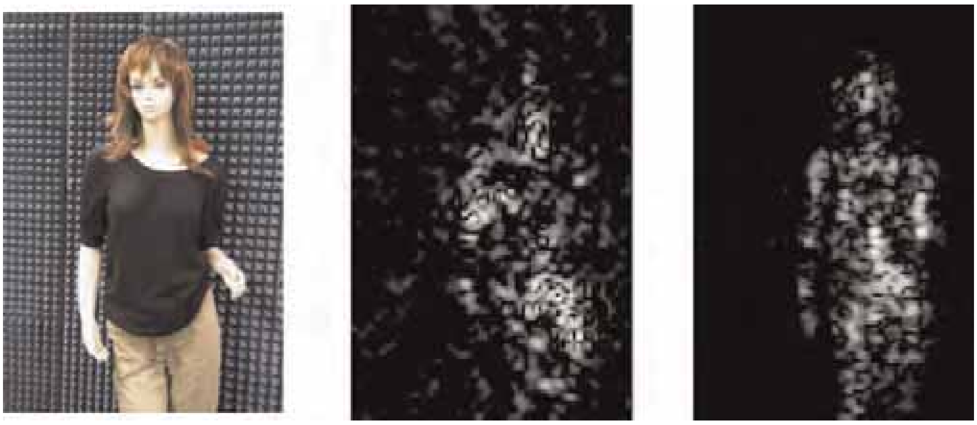

Figure 7 *Comparison of images taken by a system with separate transmitter and receiver antennas (middle) and current prototype (system with shared antenna between transmitter and receiver) (right), left is the optical image*

that sharing a single antenna, or the ESRA in this case, between the transmitter and receiver was effective in improving the image quality.

Then, to test the 3D imaging capability, a scene to be imaged was set as shown in Fig. 8*a*. Two life-size plastic mannequins and a chair were put in a laboratory. Panels of radio absorber were also placed around the scene for clarity of the test environment. The two mannequins were about 2.5 m from the prototype system. Note that the depth of the scene was not so large that the focus was fixed at 2.5 m in this case. A perspective view of the 3D profile is shown in Fig. 8*b*. The two mannequins can be clearly seen, and the backrest of the chair, a post, and some parts of the wall are also displayed. The viewpoint can be altered easily, so the position relationships between the objects in the scene can be grasped easily and intuitively.

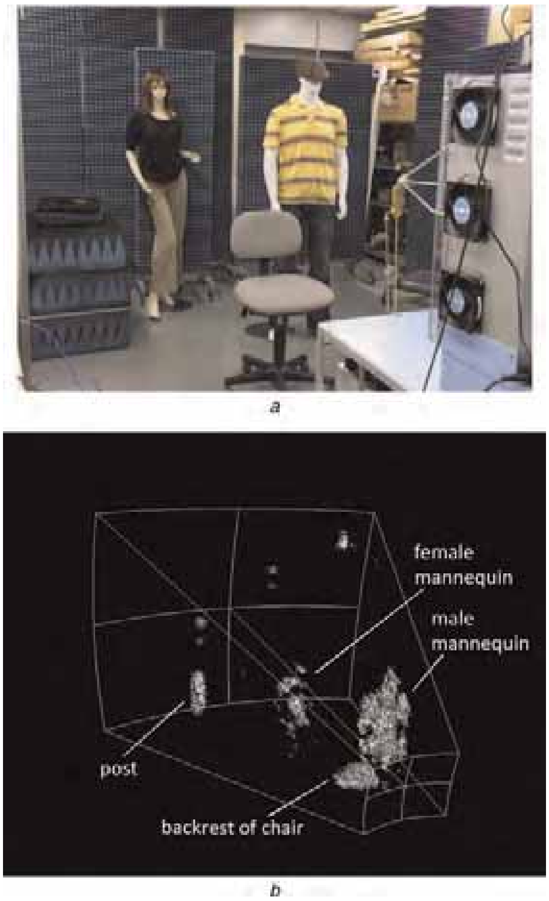

Figure 8 *Testing 3D imaging capability*

a Scene setup
b Perspective view of 3D profile taken by prototype

In addition, to investigate the detection of slight movement, a person came into the scene and stood still with raised hands next to the male mannequin. The resulting image with the movement detection feature is shown in Fig. 9, where the red voxels indicate voxels where a phase or amplitude change was detected. As opposed to the mannequins, the person was actually moving slightly even though he was standing still. Thus, the experiments proved that active millimetre-wave imaging can detect

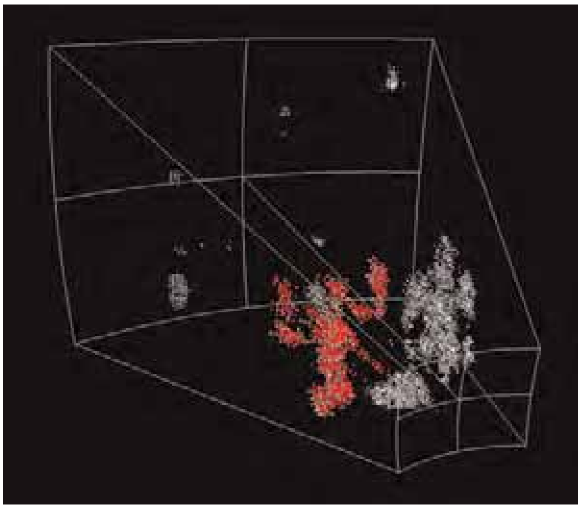

Figure 9 *Real human can be distinguished even if he/she is standing still in the same way as mannequins*

Red voxels indicate places where phase or amplitude change was detected

© The Institution of Engineering and Technology 2011

slight movements of humans, which may be advantageous in some scenarios.

4 Conclusion

A prototype active millimetre-wave imaging system using the 60 GHz band was developed to explore the possibility of applying millimetre-wave active imaging technology to broadcasting. The system has an electronic scanning reflectarray antenna (ESRA) to achieve high frame rate and variable focusing. The ESRA is shared between the transmitter and receiver antenna through simple circuit architecture to improve the image quality. Taking advantage of active millimetre-wave imaging, it acquires profiles of the scene along the range dimension as well as along the azimuth and elevation dimensions, such that a 3D profile of the scene is obtained. In addition, a feature detecting slight movement of objects, by analysing the phases of the received signals, was added.

Experimental results demonstrated that the prototype works as expected. In particular, it is expected that presenting a 3D profile and movement detection results could significantly help users to recognise the scene being imaged. More experiments assuming various scenarios are required for further evaluation and improvement.

5 References

[1] APPLEBY R., WALLACE H.B.: 'Standoff detection of weapons and contraband in the 100 GHz to 1 THz Region', *IEEE Trans. Antennas Propag.*, 2007, **55**, (11), pp. 2944–2956

[2] APPLEBY R., ANDERTON R.N., PRICE S., *ET AL.*: 'Mechanically scanned real time passive millimetre wave imaging at 94 GHz', *Proc. SPIE*, 2003, **5077**, pp. 1–6

[3] SHEEN D.M., MCMAKIN D.L., HALL T.E.: 'Three-dimensional millimeter-wave imaging for concealed weapon detection', *IEEE Trans. Microw. Theory Tech.*, 2001, **49**, (9), pp. 1581–1592

[4] KOMAROV I.V., SMOLSKIY S.M.: 'Fundamentals of short-range FM radar' (Artech House, 2003)

[5] HUANG J., ENCHINAR J.A.: 'Reflectarray antennas' (John-Wiley & Sons, Inc., 2008)

[6] KAMODA H., TSUMOCHI J., KUKI T.: 'Millimeter-wave TV camera using electronically reconfigureurable reflectarray antenna'. Proc. IEEJ Electronics, Information and Systems Society Conf., 2010, pp. 650–654

© The Institution of Engineering and Technology 2011

Selected content from IBC2011

ISSN 2041-5923

Getting machines to watch 3D for you

M. Knee

Snell, Hartman House, Danehill, Lower Earley, Reading, Berkshire RG6 4PB, UK
E-mail: mike.knee@snellgroup.com

Abstract: The advantages of automatic monitoring of multiple television channels are well known. There are just not enough eyeballs for human operators to see what is going on. With the advent of stereoscopic 3D in mainstream television production and distribution, the benefits of automatic monitoring are even greater, as 3D viewing is even less conducive to manual monitoring. This paper gives a comprehensive introduction to a wide range of automatic monitoring possibilities for 3D video. There are significant algorithmic challenges involved in some of these tasks, often involving careful high-level analysis of picture content. Further challenges arise from the need for monitoring to be robust to typical processing undergone by video signals in a broadcast chain. In this paper, the authors present the results of algorithm research into some high-level automatic 3D monitoring challenges, which can help to save human eyeballs to enjoy the 3D experience.

1 Introduction

Stereoscopic 3D is now a well established and growing part of the television broadcasting scene. Viewers at home are happy to put on 3D glasses to enjoy a new dimension to their entertainment – or maybe they are lucky enough to have an autostereoscopic display.

Running a multi-channel TV broadcast installation brings new headaches when 3D is involved. Live monitoring of dozens of TV channels is difficult enough. Over the years several manufacturers have developed automated monitoring solutions covering a whole range of tasks of increasing complexity. Examples are: detecting loss of picture or sound, lip-sync measurement and programme verification using fingerprint technology, and estimating compression picture quality by looking at blocking artefacts. With the advent of 3D, there is literally a new dimension of monitoring tasks, because we have to check not only the integrity of individual video signals but also the correct relationship between the left and right video signals in a stereo pair. In addition, manual monitoring of 3D is more difficult than 2D because the operator would need either to wear glasses or accept the limitations of autostereoscopic displays. For these reasons, there is a burgeoning interest and market in automatic monitoring of 3D television.

This paper first provides an overview of 3D monitoring problems. Subsequent Sections deal with several of these problems in detail and discuss how they might be solved. We begin with low-level and relatively simple format detection tasks and move on to analysis of depth and disparity, particularly with a view to reducing eye strain problems. Finally, two interesting examples of more algorithmically challenging tasks are presented.

Throughout this paper, the term '3D' implies stereoscopic 3D, though some of the techniques discussed may also be applied to more advanced multiview 3D representations.

2 Overview of 3D monitoring

2.1 Analysis and correction

This paper is about monitoring or analysis of 3D video. One of the purposes of automatic analysis is to provide information to enable correction of any problems encountered. The techniques for correction are beyond the scope of this paper, though it is important to point out that correction of an upstream problem may be necessary before monitoring of further aspects can be carried out.

2.2 Metadata

The correct use of metadata, for example to identify left and right channels or to signal how they are packed into a single container, can in theory remove the need for some analysis of

© The Institution of Engineering and Technology 2011

signal essence. However, metadata for 3D is not yet fully standardised, and even when it is there will still be cases of incorrect usage, so there will always be a place for techniques that avoid the requirement for metadata. Of course, the results of measurements performed at any point in the processing chain may in their turn be passed on downstream as metadata.

2.3 Format detection

The first task when faced with a single video signal carrying a stereoscopic pair is to identify the format by which the two channels are packed into one container. For some formats this is an easy task, but there are some problems when the granularity of the packing is finer. We shall deal with this problem in detail in the next main Section of this paper.

2.4 Matching left and right images

Having unpacked the signal into left and right channels, the next task is to check whether the two channels are correctly matched, particularly as regards timing, grey scale and colour balance. Grey scale and colour balance can be aligned using histogramming techniques. Relative timing can be measured using fingerprinting techniques similar to those used for lip-sync measurement. This subject is not dealt with in detail in this paper, but it is important to note that a timing mismatch will not only be detrimental to the 3D viewing experience but will also have an adverse effect on downstream analysis, particularly of 3D depth. Relative timing is thus a good example of the need to correct a problem before further analysis can reliably be performed.

2.5 Depth or disparity analysis

A more algorithmically challenging analysis task is to measure the 3D depth across the picture, which is directly related, via screen size and resolution, to disparity or relative displacement between the left and right representations of objects in the scene. Horizontal disparity that is outside a certain range, as well as undue vertical disparity, are known to cause significant problems of eye strain for some viewers [1, 2]. Disparity analysis is also important for checking the overall relative geometric alignment of the two images.

2.6 Higher level analysis

Finally, we shall look at two examples of detection tasks which require a higher level of analysis. The first is deceptively simple to state: can we tell whether the left and right channels have been inadvertently swapped? The second is: can we tell whether the 3D pair has come from a simple 2D to 3D converter? Ultimately, 3D analysis can extend to detecting or measuring any process that has been carried out on 3D signals, either with a view to improving, modifying or reversing the process, or simply in order to report or record what has been done.

3 Format detection

There are many ways in which left and right signals may be packed into a single video channel. These include left–right or top–bottom juxtaposition (with or without reflection of one of the channels), line interleaved, column interleaved, checkerboard and frame interleaved formats. For the purposes of automatic detection, these formats may be classified into two groups. Left/right and top/bottom formats are 'loose packed' because the two pictures are physically quite separate. The remaining formats are 'close packed' because corresponding left and right channel pixels are close together in space or time.

3.1 Loose packing

Loose packed formats are quite easy to detect. One way is to carry out a trial unpacking with an assumed format and then detect whether the two resulting images are sufficiently similar to be a stereoscopic pair. Also if the two images turn out to be identical, we may conclude that a 2D image is being transported in a 3D container; this is a simple case of disparity estimation in which we look for zero disparity across the image. Fig. 1 shows the left–right differences for a small area of a picture when each of four possible trial formats is used to unpack each of four possible actual formats. Where the correct format has been used, the left–right difference contains only edge information arising from disparity.

We can summarise the detection of loose-packed formats by saying that we exploit the relative similarity of the left and right images when compared with unrelated, distant parts of the picture.

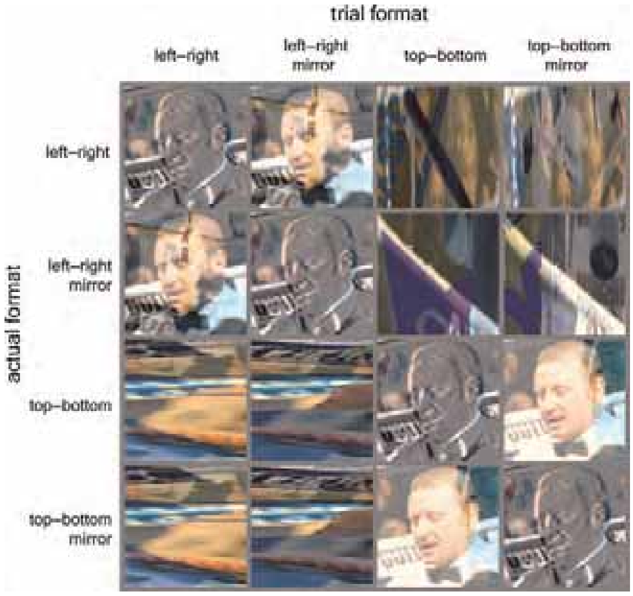

Figure 1 *Detection of loose packed formats*

© The Institution of Engineering and Technology 2011

3.2 Close packing

Close packed formats present more of a problem because the packed image looks increasingly like a single 2D image as the amount of 3D content in the scene decreases. So simply carrying out trial unpackings will often give a positive result, even if the wrong format is being tried. If there is significant 3D content, the detection becomes easier because a picture wrongly unpacked will look increasingly less like a pair of plausible images. The left half of Fig. 2 shows a small part of the left image for some different combinations of packing and unpacking formats, and the right half shows the combined energy of horizontal and vertical high pass filtered versions of those outputs. The energy is clearly significantly lower when the correct unpacking format has been used.

We can summarise the detection of these close-packed formats by saying that we exploit the relative difference of the left and right images when compared with adjacent pixels or lines.

Temporal interleaving presents further difficulties because there is a higher chance that motion can be confused with left–right disparity. This could be overcome using motion compensated highpass filtering, though care would have to be taken to use information from a single channel (albeit subsampled) for the purposes of motion estimation.

4 Depth or disparity analysis

One of the most important monitoring or analysis tasks in stereoscopic 3D is to measure the perceived depth of the various objects in the scene. Perceived depth is a function of disparity (the horizontal distance between left and right representations of the object, measured in pixels), display size and resolution, and viewing distance. In the context of signal monitoring, we can only measure disparity and then relate it to perceived depth for different display configurations.

Disparity measurement is useful for many monitoring purposes, the most important being to provide a warning if the viewer is likely to suffer eye strain. Other reasons for measuring disparity are to verify that the sequence really is 3D rather than just being 2D in a 3D container, to detect and correct for global geometric distortions between the two channels, and to assist in the insertion of captions or subtitles at suitable depths [3].

4.1 Eye strain

Eye strain can occur in 3D viewing when disparity exceeds certain limits – particularly if the eyes are being encouraged to diverge, an unnatural action. The limits depend on display size but it is also useful to measure how often and for how long extreme disparity values are observed, and possibly to identify where in the scene the extremes are occurring.

4.2 Disparity measurement

One class of disparity measurement methods involves performing a local correlation between the left and right images to generate a sparse disparity map. This approach is ideal for looking at the behaviour of different objects in the scene and for determining to what extent limits have been exceeded. Other methods seek to generate a dense disparity map, in which every pixel has an associated disparity value, or possibly an occlusion indicator if there is no corresponding point in the other picture. This approach would be necessary if the measurement were being used to drive post-processing, for example to change the effective camera spacing. Finally, for some applications an approximate, region-based approach to disparity measurement might be sufficient, for example to gather statistics about typical depth ranges used across a programme, or to drive a global spatial transform to correct for camera misalignment.

4.3 Vertical disparity

The impression of depth is conveyed by introducing horizontal disparity. If there is any vertical disparity

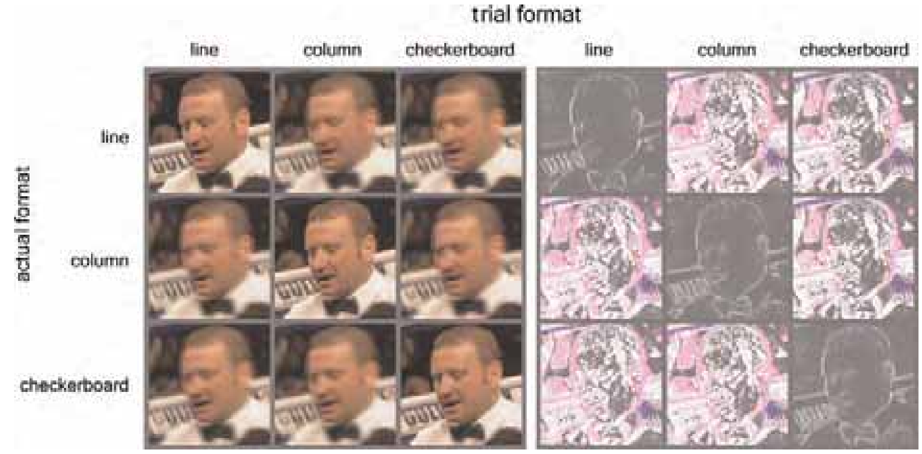

Figure 2 *Detection of close packed formats*

© The Institution of Engineering and Technology 2011

present, it should be detected and corrected, both because it can be very disturbing to the eyes, and because it can interfere with correct measurement of horizontal disparity. Of course, horizontal and vertical disparity can be measured jointly using conventional motion estimation methods. However, it would be preferable to exploit the constraints arising from stereoscopy. For example, we would expect vertical disparity to be a combination of two components: one directly related to horizontal disparity, such as might arise from a vertical displacement between the cameras, and one which fits a simple global model, such as might arise from different zoom factors or axis directions between the cameras.

4.4 Disparity monitoring display

Fig. 3 shows an example of a monitoring display that provides information about the distribution of disparity in various ways, including a left–right difference, a disparity histogram, an indication of vertical disparity and a colour coded warning of the possibility of eye strain from near and far objects for different display sizes. Such a tool makes good use of automatic analysis coupled with an operator's skill in interpreting the results.

4.5 Dense disparity maps

As a result of the difficulty and the usefulness of measuring dense disparity maps, there is some interest in standardising a format for dense disparity map metadata. For example, SMPTE has recently begun such an activity [4].

5 Higher level analysis

In this final Section we consider two examples of more challenging 3D analysis problems.

5.1 Left–right swap detection

Many people viewing 3D demonstrations have encountered the situation where the left and right images have been inadvertently swapped over. The result is very disturbing, but it is not always obvious even to a human observer what is wrong. It would be useful to be able to detect the swap automatically, but this turns out to be quite a difficult problem. Measurement of a disparity map is a good starting point, but a correctly arranged 3D pair will typically exhibit both negative disparity values for objects intended to be seen in front of the screen and positive values for objects behind the screen. So a simple analysis of the histogram of disparity values, for example, will not be enough.

One approach that works with reasonable reliability is based on the spatial distribution of disparity values. We observe that for most scenes objects at the centre and bottom of the screen are generally nearer than objects at the top and sides. Fig. 4 shows the spatial disparity distribution measured over a set of varied clips comprising 6000 frames.

A possible left–right detection algorithm is to correlate measured disparity with the above template. A positive correlation indicates that the assumed left–right configuration is correct, while a negative correlation indicates that it is reversed.

Fig. 5 shows the results of such an algorithm on 38,000 frames of (correctly ordered) 3D material. The blue line shows a 10 frame rolling average and the red line a 1000 frame rolling average of correlation coefficients between measured disparity and the template.

Whenever the graph is positive, the algorithm is giving a correct result. The last third of the material is professionally

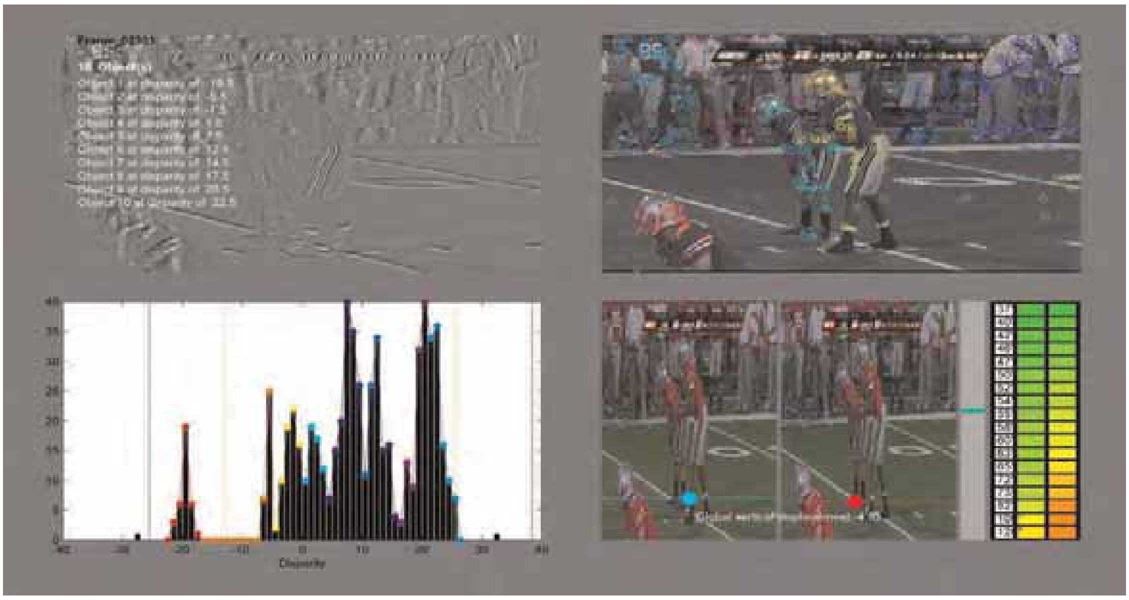

Figure 3 *Example of a disparity monitoring display*

© The Institution of Engineering and Technology 2011

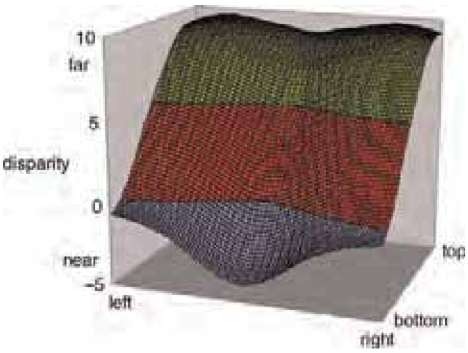

Figure 4 *Spatial disparity distribution*

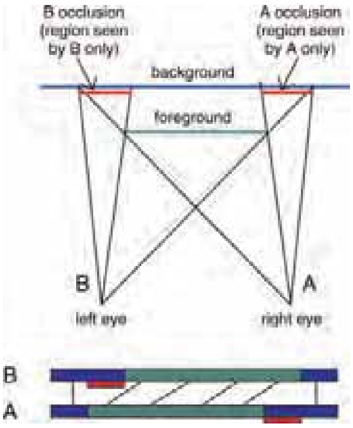
B occlusion (region seen by B only)
A occlusion (region seen by A only)
background
foreground
B
A
left eye
right eye
B
A

Figure 6 *Use of occlusions in left–right swap detection*

produced, well-behaved 3D material whereas the first two-thirds consists of test sequences of varying quality. Clearly, there is always some material that will defeat the algorithm, but on 'normal' material it is quite reliable.

A potentially more reliable method of left–right detection is based on the observation that closer objects are expected to occlude objects that are further away. A dense disparity estimator will usually have some kind of confidence output which indicates whether a pixel or region in one view has no equivalent in the other view and is therefore an occluded background region.

As shown in Fig. 6, we would expect occluded regions to extend to the left of transitions in the left-eye view and to the right in the right-eye view. The bottom part of the diagram shows where the transitions between foreground (green) and background (blue) are observed to be in relation to occlusions (red) in the two views. This observation allows us to determine automatically, on a statistical basis, which view is the left-eye view and which is the right-eye view. This approach is potentially more reliable than the method based on spatial disparity distribution, but it does depend on accurate dense disparity estimation including reliable location of occlusions.

Reliable analysis of the local relationship between depth and occlusions may be employed for other high-level monitoring tasks, for example to provide a warning that captions might have been inserted at an inappropriate location or depth relative to the other objects in the scene.

5.2 2D to 3D conversion detection

Our final example concerns the automatic detection of automatic 2D to 3D conversion. Concern is sometimes expressed that, in the rush to deliver as much 3D content as possible, content providers may resort to the use of 2D to 3D conversion. There is a great deal of interest in this field and some examples of automatic conversion are impressive, but there remains a concern that over-use of simple conversion algorithms may undermine the appeal of 3DTV. For example, Sky in the UK '… has stated it will not accept any 2D to 3D conversions for any content submitted for Sky 3D' [5]. It would therefore be desirable when monitoring 3D content to detect the possibility that a converter has been used.

One common technique in simple 2D to 3D conversion is the use of a fixed spatial disparity profile; for example the bottom and centre of the picture are made to appear closer than the top and sides, much as shown by Fig. 4 above. Another technique is to introduce delay between two versions of the same moving sequence to give an impression of depth. This can work because a 3D camera rig tracking

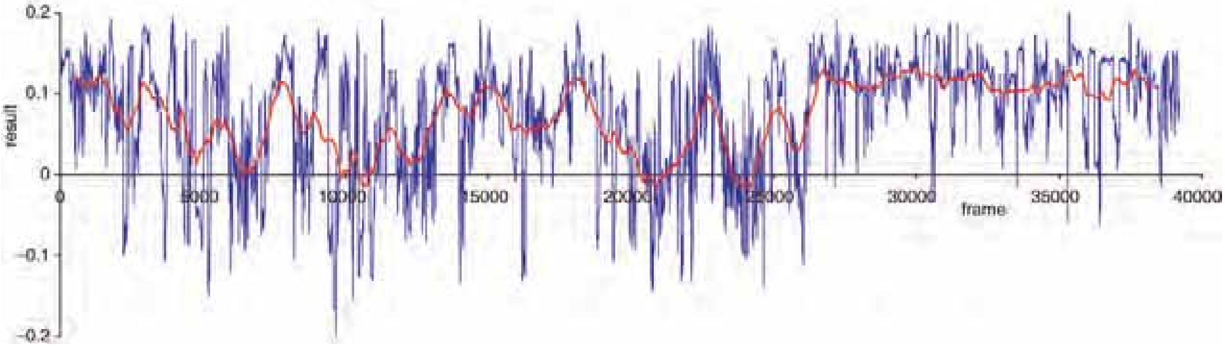

Figure 5 *Performance of left–right swap detection algorithm based on disparity distribution*

© The Institution of Engineering and Technology 2011

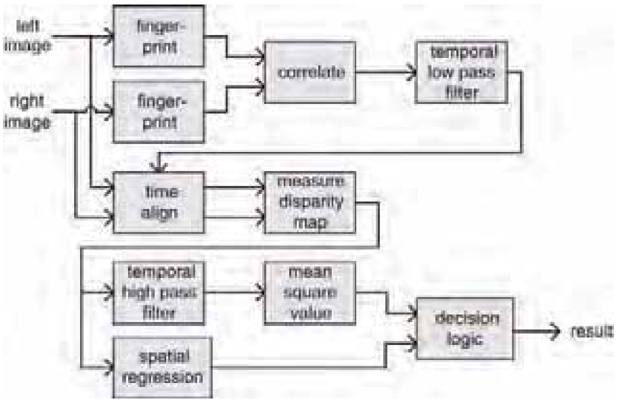

Figure 7 *Block diagram of automatic 2D to 3D conversion detector*

across a static scene will in fact generate two streams separated by a delay which corresponds to the time taken for the camera to move by the eye spacing distance.

The algorithm illustrated in Fig. 7 detects the use of either or both of these techniques, to give a warning that a 2D to 3D converter might have been used.

Fingerprints are calculated separately on the left and right input picture signals. These could be as simple as the average luminance value over each frame, an average over each of a few regions, or any measure which when applied to correctly co-timed left and right signals would be expected to be similar to each other.

A correlation process is then applied to the two fingerprint signals to produce an estimated temporal offset between the input channels. This estimated offset is applied to a temporal low pass filter, which may for example be designed to detect piecewise constant inputs. The filtered temporal offset value is used to control a temporal alignment process on the left and right images; this would be done by applying a delay to one or other of the two inputs.

A disparity map between the temporally aligned left and right images is then calculated, producing a number of disparity values across the picture. A temporal highpass filter is applied to the disparity values, thereby looking for variation in time of the disparity observed in each part of the picture. The mean square value, or other average energy value, of the highpass filter output is calculated. In parallel, a spatial regression process is applied to the disparity map to see if the map fits a fixed spatial model. A low mean square output from the temporal high pass filter, or a close correlation to a fixed spatial model, both provide evidence for a final decision that simple 2D to 3D conversion might have been performed.

With automatic detection such as this, one can envisage a game of ‘cat and mouse’ whereby detection algorithms have to become ever more sophisticated in order to keep up with the increasing complexity of automatic 2D to 3D converters.

6 Conclusions

In this paper we have described several techniques for the automatic monitoring of stereoscopic 3D video signals: format detection, disparity monitoring, left–right swap detection and automatic detection of the use of 2D to 3D conversion. We have shown that there is a great deal of scope to ‘get machines to watch 3D for us’ so that humans can concentrate on delivering and watching 3D content. Snell Ltd is active in developing and implementing the algorithms described here for monitoring and correction of 3D video across its product range.

7 Acknowledgments

The author would like to thank the Directors of Snell Ltd for their permission to publish this paper, and his Snell R&D Algorithms team colleagues, particularly Dr Roberta Piroddi, for their valuable contributions, suggestions and support.

8 References

[1] LAMBOOIJ M.T.M., IJSSELSTEIJN W.A., HEYNDERICKX I.: ‘Visual discomfort in stereoscopic displays: a review’, *Proc. SPIE*, 2007, **6490**, article id 64900I

[2] POCKETT L., SALMIMAA M., PÖLÖNEN M., HÄKKINEN J.: ‘The impact of barrel distortion on perception of stereoscopic scenes’, See http://s3.amazonaws.com/publicationslist.org/data/jukka.hakkinen/ref-69/L_Pockett_SID2010.pdf

[3] OSTERMANN R.: ‘3D subtitling’. Proc. IBC, 2010

[4] SYMES P.: ‘SMPTE. What’s happening at a standards organization near you’, See http://provideocoalition.com/index.php/awilt/story/hpa_tech_retreat_2011_day_4/

[5] PENNINGTON A.: ‘Sky bans 2D to 3D conversions’. TVB Europe, 19 March 2010, see http://www.tvbeurope.com/main-content/full/sky-bans-2d-to-3d-conversions

© The Institution of Engineering and Technology 2011

Selected content from IBC2011

ISSN 2041-5923

Automatic real-time HD defogging system

J.P. Oakley[1,2] *Z. Marceta*[3]

[1]*Dmist Research Ltd., Unit 3 Rugby Park, Battersea Road, Stockport, SK4 3EB, UK*
[2]*The University of Manchester, Manchester, M13 9PL, UK*
[3]*RT-RK Computer Based Systems LLC, Fruskogorska 11, 2100 Novi Sad, Serbia*
E-mail: j.oakley@manchester.ac.uk

Abstract: This paper describes the automatic mitigation of airlight noise (defogging) in a new in-line image processing system. Solutions to two key technical challenges are presented. The first challenge is the design of an algorithm to produce 'maps' of airlight in RGB space using information derived from sample frames. The second challenge is to process the HD stream on a pixel-by-pixel basis with low latency by appropriate subtraction and re-scaling. Although the video processing is relatively simple, it is necessarily performed in RGB space and so colour conversions are required to translate from and to the YUV representation used in transmission, leading to a significant computational requirement. This requirement is met by an asynchronous dual-processor architecture that allows sample frames to be downloaded for airlight analysis with concurrent high-speed pixel processing. Test results show effective enhancement of degraded images with no distortion of clear images and no requirement for the user to adjust settings for different conditions. The latency for 1080i/50 streams is 71 μs.

1 Introduction

Significant loss of image quality can arise in adverse atmospheric conditions such as rain, drizzle, smoke and fog. This is owing to light scattering from particles between the camera and the subject, generating what is often called 'airlight'.

Under clear conditions, the only light entering a camera is that directly reflected by objects in the field of view. If haze, fog, drizzle, rain or light smoke is present then some of the light originating from the primary light source (normally the sun, but could also be an artificial light source) is scattered so that it enters the camera. This is known as the 'airlight' and the effect is illustrated in Fig. 1. The resultant image, I(x, y), produced by the camera is essentially a sum of two components: the scene component S(x, y) and an airlight component A(x, y):

$$I(x, y) = S(x, y) + A(x, y) \tag{1}$$

The intensity of the airlight is a function of the size and composition of scattering particles, the concentration of particles, the distance between subject and camera and the angle of illumination.

All these parameters are subject to change – some over the image area, and some over time. The amount of scattering is also dependent on the wavelength of the light (e.g. sometimes greater for blue than red wavelengths). The scene component S(x, y) is also attenuated by the atmosphere; a process known as extinction. The combined effect of these scattering phenomena is to reduce image contrast.

Airlight degradation is an important problem in outside broadcast, particularly with HD services as viewers pay a premium for high picture quality. The best current solution is manual adjustment of black level at the video editing desk using 'Proc Amp' controls. However the colour can easily be distorted and different parts of the image can vary in brightness. It is also a demanding task when the camera is tracking a moving subject.

This paper describes the automatic mitigation of airlight noise (defogging) in the new in-line image processing system called ClearVue. An example is shown in Fig. 2. Only the left-hand part of the image is processed to allow comparison.

Two key technical challenges addressed in the design of ClearVue are presented. The first challenge is the design of a reliable algorithm to produce 'maps' of airlight in RGB space

© The Institution of Engineering and Technology 2011

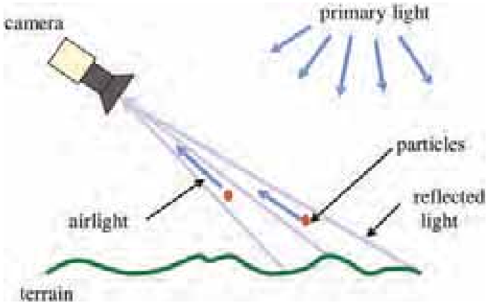

Figure 1 *Effect of haze*

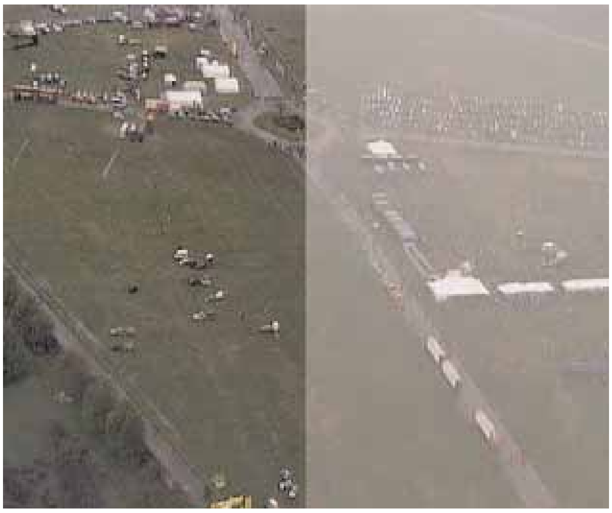

Figure 2 *Left part enhanced*

using information derived from sample frames. The second challenge is to process the HD stream with low latency by appropriate subtraction and re-scaling. In this paper the theoretical background for the image enhancement is described. An overview of the implementation via an asynchronous dual-processor architecture is then given. Some illustrative results are presented followed by discussion.

2 Background

2.1 Previous approaches

Mitigation of this type of atmospheric degradation can be effected in various ways. The best-known image enhancement tools are based on histogram equalisation. Most of these programs will provide some improvement in image quality when applied to atmosphere-degraded images. However the time required for the enhancement computations introduces a delay, known as latency, between input and output. Latency is an important issue in outside broadcast. Also the previous enhancement algorithms distort clear images. More recently, specialised algorithms have been reported to mitigate atmospheric degradation [1–5]. Such algorithms are idempotent in the sense that they correct a specific defect in the image. If no atmospheric degradation is present then they will introduce no changes in the image.

The basic idea is to invert (1) to recover S(x, y) from I(x, y). If the airlight distribution A(x, y) can be estimated by some means then the image may be recovered by simply subtracting A(x, y) from I(x, y), followed by appropriate rescaling.

The reported algorithms differ in how the airlight component A(x, y) is estimated. In (1) the fact that the airlight varies with range is exploited using non-linear regression to produce an estimate for A(x, y). In cases where the range does not vary significantly across the image this approach runs into difficulties. Narasimhan and Nayar [2, 4] describe a method for producing airlight estimates on the assumption that the range (and hence the airlight) is piecewise constant in the images. Again this method runs into difficulties in applications where the assumption is not valid. Oakley and Bu [3] describe a more general method based on minimising a cost function. This latter technique can work in the widest range of conditions and so is preferred here. An outline of this method is given below.

2.2 Cost function approach

The coefficient of variation (CV) is defined as the ratio of the local standard deviation of the pixel intensities to the local mean.

The assumption in the Oakley–Bu method is that the statistics of a clear image are, to a first approximation, stationary. An image typically contains some dark objects and some brighter regions. The assumption is that the CVs in the bright and dark regions should be similar. This is reasonable for natural scenes since differences in illumination generate image regions with different lightness but similar CVs. In foggy conditions the CVs differ considerably for light and dark regions. This is illustrated by the simulation shown in Fig. 3. Two synthetic images are shown; the first represents a clear image with a constant CV. An estimate for the CV is calculated from the equation displayed in which p_k is the value of the image at pixel position k and $\overline{p_k}$ is the output of a spatial low-pass filter at pixel position k. The second image is transformed using (1) to represent the foggy case. Plots of the CV for one selected line are shown for both clear and foggy cases. It can be seen that the clear image has a relatively uniform CV, although it is subject to statistical fluctuation. The foggy image shows greater variation in the CV, with darker regions showing lower values. This is a fundamental difference that can be detected by appropriate statistical analysis and this is the basis of the Oakley–Bu method.

The Oakley–Bu cost function is:

$$S(A) = \frac{1}{k}\sum_{k=0}^{k<K}\left(\frac{p_k - \overline{p}_k}{\overline{p}_k - A_k}\right)^2 \cdot \exp\frac{1}{k}\sum_{k=0}^{k<K}(\ln(\overline{p}_k - A_k)^2) \qquad (2)$$

The airlight values $\{A_k\}$ are chosen to minimise the value of this function. Even with a good estimate for $\{A_k\}$ the CV will still vary significantly in different parts of the images.

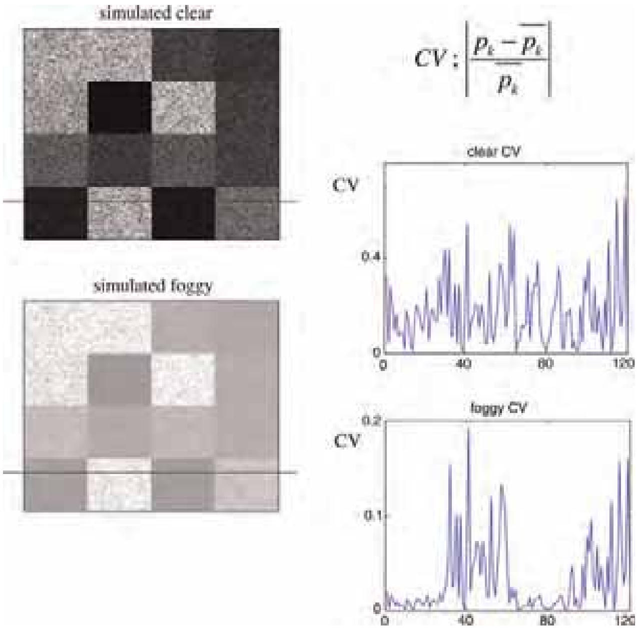

Figure 3 *CV variation in clear and foggy images*

However the spread of values of the CV is minimised in a specific sense (it can be shown that the Oakley–Bu cost function is equivalent to minimising the Theil index T0, a well-known metric for variability used in the analysis of economic inequality). A can be represented as a parametric function, as in (1), in a global minimisation or as a smooth non-parametric function, in which case some kind of iterative local minimisation is required. The latter approach is used in the ClearVue system to give the greatest possible flexibility in application. The airlight estimation algorithm is coded in C++ and implemented on a conventional IA32 processor.

3 Implementation

Once airlight has been estimated the required enhancement computation is a pixel-by-pixel subtraction and scaling. The level of airlight in general varies with wavelength and hence is different for the red, green and blue channels. For this reason the processing is performed in RGB colour space. If the input pixel is (x_r, x_g, x_b) and the output pixel (y_r, y_g, y_b), then the required transformation is

$$\begin{aligned} y_r &= m_r\,(x_r - A_r) \\ y_g &= m_g\,(x_g - A_g) \\ y_b &= m_b\,(x_b - A_b) \end{aligned} \qquad (3)$$

where m_r, m_g, m_b , A_r, A_g and A_b represent scaling and airlight (offset) parameters. The required gain and offset parameters may vary for different parts of the image since the extent of the degradation will depend on range. Since the actual video processing is very simple, i.e. subtraction and scaling according to (3), it is advantageous to separate the relatively complicated statistical analysis algorithm from the video processing pipeline. In this way such enhancement can be applied to a high-definition video stream in real-time [6] whilst achieving low latency (in the order of microseconds).

For medium-volume applications the video pipeline could be implemented either using a field programmable gate array (FPGA) or a digital signal processor (DSP). The DSP route was chosen for ClearVue, mainly on the grounds of lower production cost. The DSP device selected is the DM642 from Texas Instruments. The assembly is mounted in a 1U enclosure as shown in Fig. 4 above. The processing architecture is shown in Fig. 5 below. The two dotted boxes show functionality implemented on a bespoke DSP printed circuit board and functionality implemented in software on a general-purpose IA32 board. The two boards are linked via a PCI connector.

Although the pixel processing is simple, for HD streams the computational requirement is such that carefully optimised DSP code is required. As the processing must be carried out in RGB space, colour conversions, both from and to YUV colour space, are required. Conventional video sources are gamma-encoded and this non-linear transformation must also be reversed prior to processing and re-applied after processing. The central task of the video processor is to implement the transformation specified by (3) using stored values of m_r, m_g, m_b, A_r, A_g and A_b. These enhancement coefficients are held in high-speed memory within the DSP system. In principle the video process can be achieved on a pixel-by-pixel basis. In practice, in a DSP implementation, it is advantageous to process up to four lines of the image at a time. This increases the latency but the values are considered acceptable for outside broadcast application. Table 1 shows the latency values for different video formats. The HD values are lower because the coding of the DSP program so highly optimised for these cases. An FPGA-based video

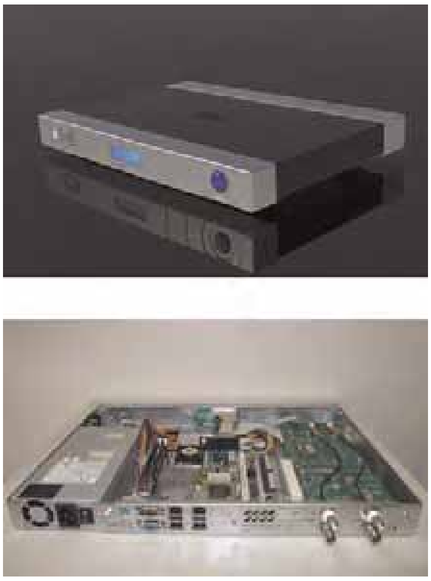

Figure 4 *ClearVue HD product*

© The Institution of Engineering and Technology 2011

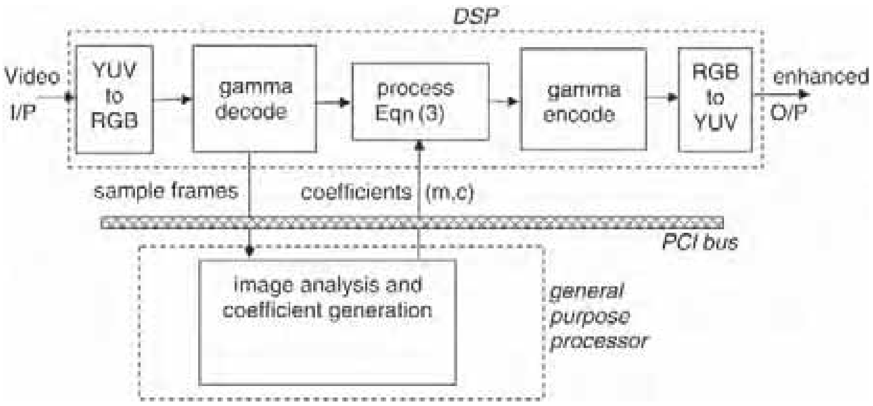

Figure 5 *Implementation architecture*

processor could be used to provide lower values of latency if required.

3.1 Operation

In operation the image analysis and video processing run as asynchronous tasks, communicating via a PCI bus. The image analysis process signals the video process when it is ready to analyse an image. An image is then transferred without affecting the ongoing video process. When the analysis is complete, new enhancement coefficients are sent to the video processor and loaded during the blanking interval. The image analysis task then requests a new frame, and so on. Although the video process operated at full video rates (50Hz for PAL and 60 Hz for NTSC), only a subset of frames, typically one in four, are used for airlight analysis. The reason for this is that the pattern of atmospheric degradation changes relatively slowly.

4 Results

The main testing methodology used with ClearVue is the processing of many hours of archive footage of different subject matter acquired under a wide variety of atmospheric conditions, followed by painstaking subjective analysis. This shows consistently high output quality with no visible distortion. Testing with live camera feeds is also used. The aim of this testing is to establish:

1. *Safety:* If there is no adverse atmospheric condition, the image should ideally not be changed at all. At worse any change should not adversely affect subjective image quality and

Table 1 Latency of enhancement process

Input	Resolution	Rate, Hz	Latency, μs
SDTV	720 × 576i	50	256
	720 × 480i	60	256
HDTV	1280 × 720p	50/60	53.33/44.44
	1920 × 1080i	50/60	71.11/59.26

2. *Effectiveness:* When adverse conditions are present the enhancement process image should be fully effective.

When the visibility is very poor the scaling effect of the transformation described by (3) increases any noise present in the image. This puts a fundamental limitation on the enhancement process since some noise is always present. The two main sources of noise are sensor noise and particle noise. Sensor noise arises mainly from shot noise caused by the discrete nature of the light detection process. Sensor noise is always present. Particle noise is caused by relatively large particles close to the lens and the level of particle noise varies greatly according to the atmospheric conditions. An extreme example would be snow. In general the ClearVue process is most effective in moderate visibility conditions where the scaling effect introduced by the enhancement does not raise noise levels to unacceptably high levels.

In some situations the atmospheric conditions can change quickly and corresponding sections of archive footage are particularly useful in testing. Fig. 6 shows image frames extracted at 0.5 s intervals under conditions of light rain which improve rapidly. The contrast for each of the five unprocessed images, defined as $(I_{max} - I_{min})/I_{mean}$, where I is image intensity or lightness, ranges from around 0.3 in image 1 to around 1.0 in image 6. The contrast for each of the unprocessed images is shown by the front bars in Fig. 7. The processed images are shown to the right in Fig. 5 and the corresponding contrast is shown by the rear bars in Fig. 7. The processed contrast is relatively stable at around 2.2. Although the subjective quality of images 1 and 2 is improved by enhancement, the enhanced images show a high degree of noise and would not be suitable for production purposes. Images 3–5, with an unprocessed contrast ratio of between 0.6 and 1.0, represent situations in which the enhancement renders the video stream usable for production. Without some kind of processing this stream could not be used. The subjective effect of this processing is that the atmospheric problem is not noticeable by the viewer. More examples of ClearVue processing can be found in [7].

© The Institution of Engineering and Technology 2011

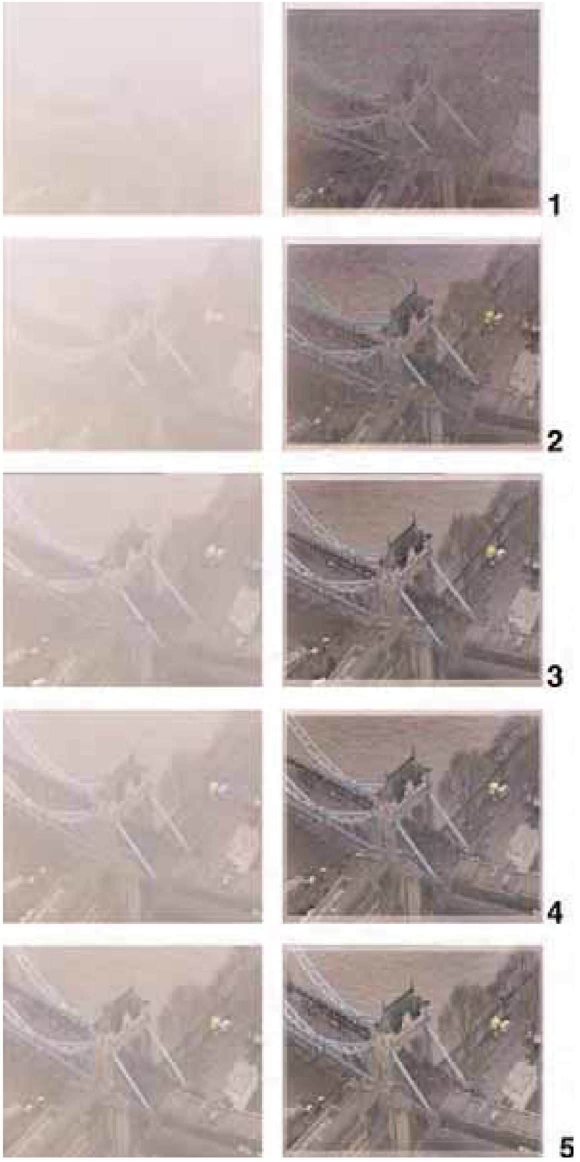

Figure 6 *Sample results (left is original and right is enhanced)*

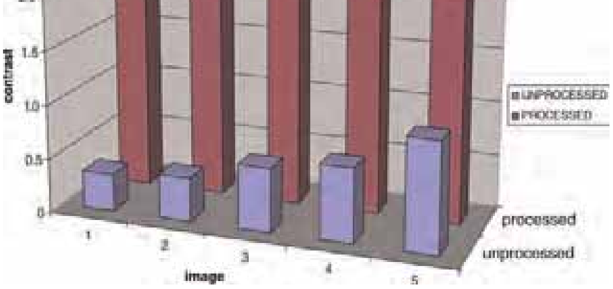

Figure 7 *Image contrast*

5 Discussion

ClearVue is the first commercially-available defogging system specifically designed for outside broadcast applications. It is effective in processing images in moderately poor visibility and restoring correct contrast and colour. The ClearVue process is completely automatic and there are no parameters to set. The system can be regarded as a kind of 'fog filter'; when no airlight is present the video stream is not altered. Viewers are generally unaware that any processing of the video has taken place unless side-by-side presentation of unprocessed video is offered.

For very poor visibility the system will improve the clarity but the output will not, in general, meet retransmission standard. The main limiting factor is the noise present in the image. In general this is owing to a combination of sensor noise (mainly shot noise) and noise introduced by particles close to the sensor.

Awareness and acceptance by the OB community will be an important milestone for ClearVue. ClearVue is currently available as a standalone add-on unit but the technology could potentially be incorporated within camera systems. The best approach will be to offer the algorithm as a set of integrated circuits and significant investment will be required to achieve this.

6 References

[1] OAKLEY J.P., SATHERLEY B.: 'Improving image quality in poor visibility conditions using a physical model for contrast degradation', *IEEE Trans. Image Process.*, 1998, **7**, (2), pp. 167–179

[2] NARASIMHAN S.G., NAYAR S.K.: 'Contrast restoration of weather degraded images', *IEEE Trans. Pattern Anal. Machine Intell.*, 2003, **25**, (6), pp. 713–724

[3] OAKLEY J.P., BU H.: 'Correction of simple contrast loss in color images', *IEEE Trans. Image Process.*, 2007, **16**, (2), pp. 511–522

[4] NAYAR S.K., NARASIMHAN S.G.: 'Vision in bad weather'. Proc. IEEE Int. Conf. on Computer Vision, 1999, Vol. 2, pp. 820–827

[5] TAN K.K., OAKLEY J.P.: 'A Physics-based approach to color image enhancement in poor visibility conditions', *J. Opt. Soc. Am. A*, 2001, **18**, (10), pp. 2460–2467

[6] ROBINSON M.J., ARMITAGE D.W., OAKLEY J.P.: 'Seeing in the mist: real time video enhancement', *Sensor Review*, 2002, **22**, (2), pp. 157–161

[7] Dmist Research Ltd., Manchester, UK. http://www.dmist.com

© The Institution of Engineering and Technology 2011

Selected content from IBC2011

ISSN 2041-5923

Combining panoramic image and 3D audio capture with conventional coverage for immersive and interactive content production

G.A. Thomas[1] *O. Schreer*[2] *B. Shirley*[3] *J. Spille*[4]

[1]BBC R&D, South Lab, BBC Centre House, 56 Wood Lane, London, W12 7SW, UK
[2]3D Video & Immersive Media Group, Image Processing Department, Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany
[3]Acoustics Research Centre, Newton Building, University of Salford, Greater Manchester, M5 4WT, UK
[4]Audio & Acoustic Lab, Technicolour, Deutsche Thomson OHG, Karl-Wiechert-Allee 74, 30625 Hannover, Germany
E-mail: graham.thomas@bbc.co.uk

Abstract: The media industry is currently being pulled in the often-opposing directions of increased realism (high resolution, stereoscopic, large screen) and personalisation (selection and control of content, availability on many devices). A capture, production and delivery system capable of supporting both these trends is being developed by a consortium of European organisations in the EU-funded FascinatE project. This paper reports on the latest developments and presents results obtained from a test shoot at a UK Premier League football match. These include the use of imagery from broadcast cameras to add detail to key areas of the panoramic scene, and the automated generation of spatial audio to match the selected view. The paper explains how a 3D laser scan of the scene can help register the cameras and microphones into a common reference frame.

1 Introduction

It is an often-expressed view that the TV industry should adopt a common video production format, which would not only be unified across the world, but also support a wide range of applications. Traditionally, the shot selection, framing and audio mix is designed to support the particular 'story' that the director is aiming to tell, and will have been produced with a particular reproduction system in mind (e.g. widescreen HD with 5.1 surround sound). Although some provisions are sometimes made to allow repurposing for other devices (such as maintaining a 4:3 'safe area' within a 16:9 frame), such content is not ideal for supporting extreme variations in viewing device, e.g. from mobile phones to ultra-high-resolution immersive projection systems with 3D audio support. Audiences increasingly expect to be able to control their experience, for example by selecting one of several suggested areas of interest, or even by freely exploring the scene themselves. Traditionally-produced content offers very limited support for such functionality. Whilst such a degree of freedom may not be appropriate for all kinds of content, it has the potential to add useful interactivity to any kind of programme where there is no obvious single 'best' shot that will satisfy all viewers.

An approach to overcoming the limitations of current production systems to help meet these requirements is the so-called 'format-agnostic' approach [1]. The main idea of this is to develop a completely new production system, which does not use fixed numbers of frames, lines and pixels, or even geometry. Such an approach requires a paradigm shift in video production, towards capturing a format-agnostic representation of the whole scene from a given viewpoint, rather than the view selected by a cameraman based on assumptions about the viewer's screen size, loudspeaker setup and interests.

The ideal format-agnostic representation of a scene would involve capturing a very wide angle view of the scene from each camera position, sampled at a sufficiently high resolution that any desired shot framing and resolution could be obtained. However, this is not only

© The Institution of Engineering and Technology 2011

impractical, but would be wasteful, as less interesting areas of the scene would be captured at the same high resolution as the key areas of interest. This leads to the concept of a 'layered' scene representation, where several cameras with different spatial resolutions and fields-of-view can be used to represent the view of the scene from a given viewpoint. The views from these cameras can be considered as providing a 'base' layer panoramic image, with 'enhancement layers' from one or more cameras more tightly-framed on key areas of interest. Other kinds of camera, such as high frame-rate or high-dynamic range, could add further layers in relevant areas. This 'layered' concept can be extended to audio capture, by using a range of microphone types to allow capture of the ambient sound field, enhanced by the use of additional microphones to capture localised sound sources at locations of interest. This allows an audio mix to be produced to match any required shot framing, in a way that can support reproduction systems ranging from mono, through 5.1, to higher-order Ambisonics (HOA) or wave field synthesis (WFS).

This paper presents some of the latest results of the EU-funded FascinatE project, which is developing a capture, delivery and reproduction system to evaluate the concepts outlined above. The project addresses several different levels of interactivity: at simplest, the production tools developed could be used to allow local or specialist broadcasters to customise and tailor coverage of live events for a specific audience. In this scenario, the users' experience will not be interactive although will be improved by being tailored to their locality and interests (for example, by showing a sporting event in a manner designed for supporters of a particular team). At the other extreme, all captured content could be delivered to the user. This would allow them to switch between a number of shot sequences selected by the director, optimised locally for their particular screen size. Users could even construct and define their own shot selection and framing, with matching audio that they could further customise, for example by adding various commentary channels.

The following Section describes the approach being taken to scene capture for both audio and video, and how a 3D laser scan of the scene can be used to register all sources in a common reference frame. This is followed by a report on a test capture carried out at a Premier League football match in October 2010, illustrating the first practical application of the ideas to acquire a data set to support the work of the project. Two specific aspects of production using the layered scene are then discussed: the use of conventional HD broadcast cameras to provide additional detail in key areas, and the rendering of the captured audio to match the chosen view of the scene. Further details of the way in which the project is handling audio may be found in [2], and a discussion of the approach being taken to the delivery network and end-user terminal is given in [3].

2 Scene capture

2.1 Video

Building on the concept of a layered scene representation, the approach taken by the FascinatE project is to make use of any available video feeds from conventional broadcast cameras, and capture additional very-wide-angle images from one or more locations co-sited with these cameras. The wide-angle capture makes use of an ultra-high-resolution omnidirectional camera – the so called OmniCam (see Fig. 1). With this system a full 180° panoramic view can be captured resulting in a total resolution of 7 × 2K pixels. Details of the system can be found at [1].

Owing to the high resolution of the captured image, fast-moving objects in the foreground become blurred, owing to the current relatively low capturing frame rate of 30 fps. Hence, in the next revision of this system, a new camera [4] will be used which overcomes this limitation. This new camera operates at 50/60 fps and moreover, is equipped with a high-quality sensor with high dynamic range, low noise, and brilliant image quality, especially for difficult lighting situations. The use of a high dynamic range camera is particularly important for panoramic imaging applications, as the field-of-view is very likely to encompass both very bright areas (such as sky) and very dark areas (such as shadows).

2.2 Audio

The FascinatE project presents a number of interesting challenges for audio capture: first the format-agnostic approach of the system requires all audio to be captured in such a way that they it be rendered across the full range of current reproduction systems, and secondly, that the audio is in a form that can be rendered to take into account the interactive control that the user will have.

Audio reproduction formats represented in the FascinatE project include stereo, HRTF generated binaural reproduction, 5.1, 7.1 surround systems, higher-order Ambisonics and wavefield synthesis.

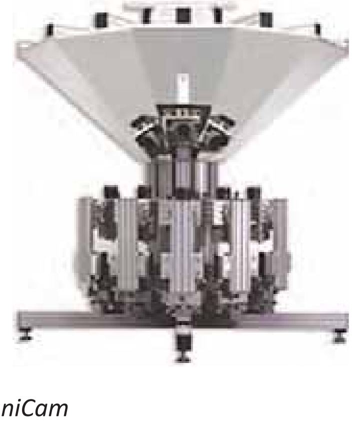

Figure 1 *OmniCam*

© The Institution of Engineering and Technology 2011

A particular challenge for audio is posed by the necessity within FascinatE to match the sound of the event to the visual effect of zooming into the picture. Although in reality the user is zooming into a 2D video, the visual effect in some cases will be that the user's position travels past objects that will move to the sides and behind the viewing position as they move out of shot. For this reason FascinatE audio must have a depth dimension that has to be mapped to the panoramic 2D video scene. For example, if while watching a football match the user zooms past the ball position to a region of interest at the opposite side of the pitch, their expectation is likely to be that the sound of the ball being kicked will move behind their new viewpoint.

To allow audio to be reproduced to match the visual appearance of the scene it is necessary not just to capture a sound field from the camera position, but instead to capture 'audio objects' with appropriate co-ordinate positions so that they may be rendered to any point around the user. The capture mechanism to allow this feature is very much dependent on the particular situation of the recording. For some events close microphone techniques at audio sources can be used to accurately generate audio objects that can be manipulated in response to user control. For other events, such as the football match described below, the situation is considerably more challenging. Further details of the way in which the project is handling audio may be found in [2].

2.3 3D scan

To register the different sensors of the FascinatE system in a common co-ordinate system, a 3D laser scanner [5] is used. This scanner allows an accurate 3D scan of a large environment such as a football stadium, including recognition of special markers. This allows the correct measurement of 3D positions of all the different sensors, such as microphones and cameras. The scanner not only provides a 3D 'point cloud' representing the scene, but also a colour image. In Fig. 2 (left), the 3D scanner is shown and on the right, the planar view of the captured colour image is presented.

In addition to directly measuring the locations of the various cameras, the 3D scan data can be used to help estimate the pan, tilt and field-of-view of the broadcast cameras, by providing an accurate depth map of features visible in the background. Computer vision techniques can then be used to identify features in the broadcast camera images and thus track the camera movement [6], for example by matching them with features visible in the OmniCam.

3 Test shoot

On 23 October 2010, the FascinatE consortium carried out the first test shoot at a live event: the UK Premier League football match Chelsea vs. Wolverhampton Wanderers, at Stamford Bridge, London. The aim of this shoot was to get a complete set of audiovisual material to research and develop the new concepts of format-agnostic production. Therefore the omnidirectional high-resolution camera system [1], the new high-dynamic range camera [4], an Eigenmike® and two Soundfield® mics were brought to London and installed on different camera platforms in the stadium (see Fig. 3). Thanks to close co-operation between BBC and their outside broadcast supplier, the consortium was able to get the recordings of four broadcast cameras, twelve shotgun microphones and several stereo microphones located around the pitch.

Various practical issues had to be overcome during the test shoot. In particular, whilst rigging the omnidirectional camera system, care had to be taken in locating it so that the views of spectators were not obstructed. Rain also posed another potential problem, as any drops of water on the mirrors or upward-facing cameras would impair the panoramic image. Luckily, the weather remained dry. After the match, a complete 3D laser scan of the stadium was captured. In this way, it was possible to accurately register all the camera and microphone positions as required for matching of visual and sound events.

It was impossible to attach microphones to the players or referee, and even techniques such as microphone arrays for localising and capturing audio sources were impractical owing to the limitations imposed by the event.

Out of necessity the FascinatE project therefore took advantage of existing recording equipment used at the

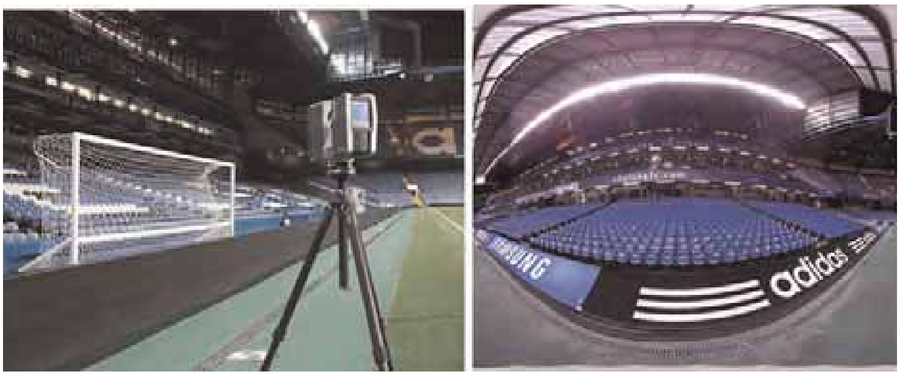

Figure 2 *3D laser scanner (left), captured planar view (right)*

© The Institution of Engineering and Technology 2011

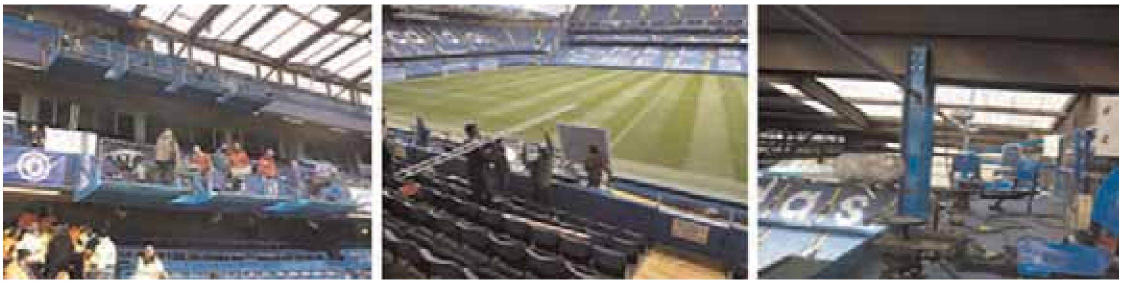

Figure 3 *Chelsea test shoot: camera platform (left), calibration of the OmniCam (middle), Soundfield$^{®}$ mics and stereo pair (right)*

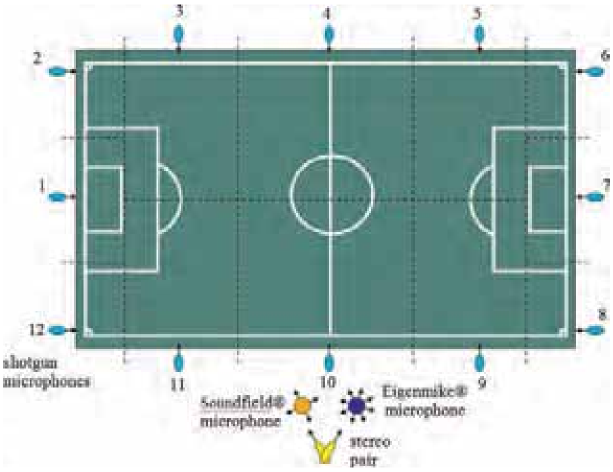

Figure 4 *Microphone positions*

stadium: 12 shotgun microphones spaced around the pitch (for on-pitch sound) and added several sound field microphones – Soundfield® microphones at either end of the half-way line and a single 32 capsule Eigenmike$^{®}$ situated close to the camera position (Fig. 4). Using these microphones a scenario has been developed whereby areas of the pitch determined by microphone placement have been defined as static audio objects that may be either active or inactive depending on automatic assessment of key audio events. This combination then allows the user to dynamically change their viewing direction and apparent location with appropriate panning effects being applied to sound sources.

The audio and video content contained all the 90 minutes of the match of which about 10 minutes (occupying about 1 Tbyte) was selected for distribution to the consortium members. From the selected clips of the omnidirectional camera, a fully stitched panorama has been produced and made available (see Fig. 5).

4 Merging of broadcast cameras into panoramic image

As discussed in the Introduction, one aim of the FascinatE project is to evaluate the 'layered scene' concept. One aspect of this is the use of the broadcast cameras to provide higher resolution to key areas of the panoramic scene. To evaluate the potential gain from this approach, tests were conducted with some of the images from the test shoot.

The OmniCam horizontal resolution is approximately 7K pixels, which covers 180 degrees – an equivalent resolution to an HD camera with a horizontal field-of-view of approximately 50 degrees. The main camera covering a football match typically has a horizontal field-of-view of around 30 degrees, although close-up cameras can go as tight as 5 degrees or less. To get the equivalent resolution of such a tight zoom from a 180 degree camera would require a horizontal image resolution of approximately 70K pixels. Using a broadcast camera to enhance resolution in areas of interest thus has the potential to increase the resolution by around a factor of 10 in each direction – well beyond what a practical omnidirectional camera could achieve. Fig. 6 shows a comparison of the resolutions.

Some initial experiments have been conducted to assess the challenges in forming a composite image from broadcast and

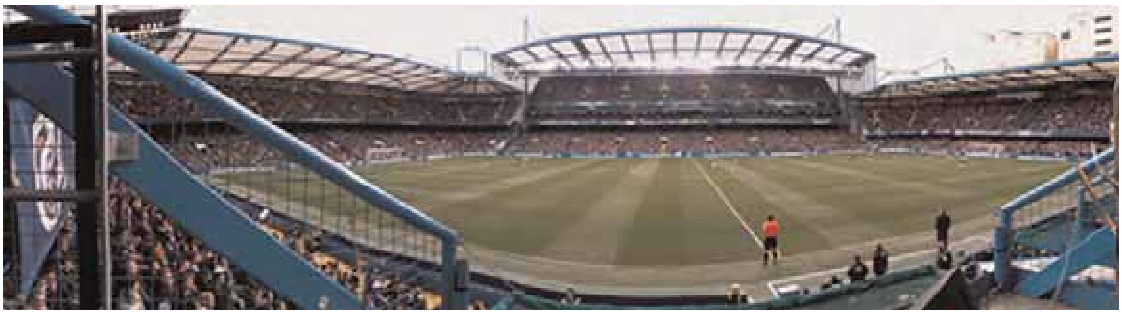

Figure 5 *Stitched panoramic view*

© The Institution of Engineering and Technology 2011

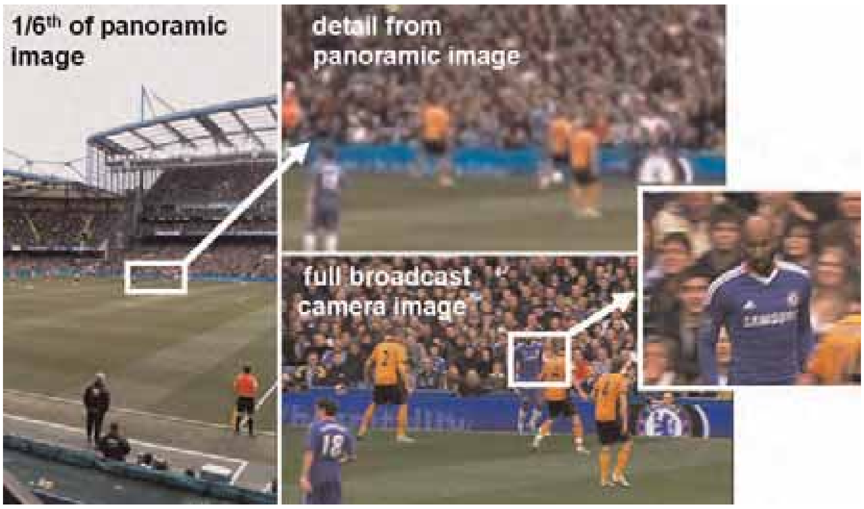

Figure 6 *Comparison between OmniCam and broadcast camera*

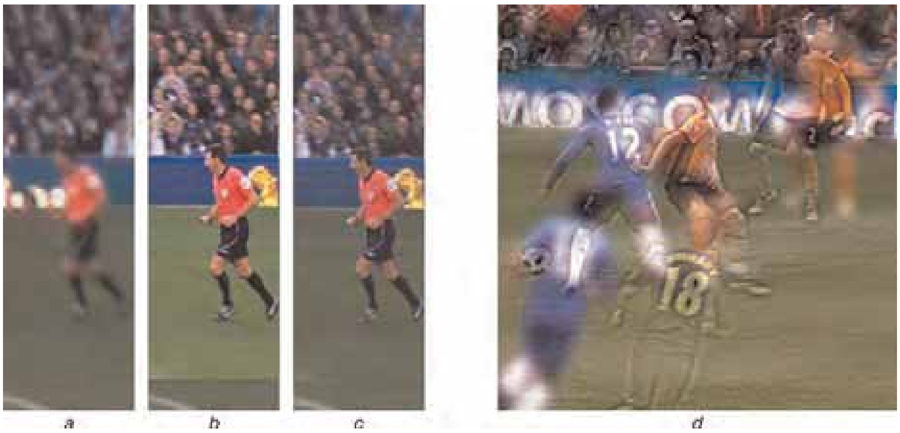

Figure 7 *Comparison between approaches for merging images*

a OmniCam image
b Direct overlay of broadcast camera image over the central part
c Overlay after colour histogram matching
d Taking high frequencies from broadcast camera

OmniCam images [7]. An issue to be overcome is mismatches in the brightness and colorimetry of the cameras. One approach that has been investigated is the use of histogram matching: the RGB histogram of each image is evaluated in the area of overlap, and a lookup table is computed to re-map the colours of one image to make the two colour histograms match. Figs. 7*a–c* shows a small part of the OmniCam image, into which a section from the broadcast camera has been overlaid. The colour mismatch is clearly visible in the central image, particularly on the grass. The colour histogram equalisation that has been applied in the right-hand image has virtually eliminated any obvious colour difference.

An alternative approach, which avoids having to correct for any level shifts, is to take the high frequency components from the broadcast camera image, and the low frequency components from the OmniCam. This guarantees that flat areas of colour will match exactly. The approach could be extended to use an adaptive filtering strategy, to ensure that detail could instead be taken from the OmniCam where this happened to give more high frequency energy (e.g. in areas of the background that suffered from motion blur in the moving broadcast camera).

Initial experiments with this approach have shown that its success depends critically on the accurate alignment of the images. In this test shoot, there was a distance of around 3 m between the OmniCam and the broadcast camera capturing the close-ups, and this resulted in significant parallax differences between the images. Fig. 7*d* shows an example of a part of a composite image, where the low spatial frequencies have been taken from the OmniCam and the high spatial frequencies from the broadcast camera. The images were aligned to match the two players near the centre. Although the detail layer from the broadcast camera correctly enhances the appearance of these players, the background and players at other depths show significant misalignment. Whilst it would be possible to apply some

© The Institution of Engineering and Technology 2011

disparity compensation to the processing, it is clear that there would be significant areas of the scene that were only visible in one of the two cameras. In this situation, it is unrealistic to expect to be able to produce a perfect merged image, and instead we are aiming to identify the best approach to producing a visually-acceptable transition between the cameras, so that a virtual zoom could be produced, starting on a wide shot from the OmniCam, and ending up with the close-up from the broadcast camera. This would meet the requirement for a user to be able to seamlessly move from viewing a wide shot to a region-of-interest covered by a broadcast camera.

5 Generation of audio to match video

One principle of FascinatE is to transmit as much information as possible to the terminal in its original format, rather than transcoding from one format into another. Therefore audio objects and sound field recordings are transmitted separately. This allows the user to interact with the content independently, for example selecting audio objects like the TV commentator and rotating the sound field depending on the viewing direction. At the terminal the sound field signal will be decoded and the audio objects will be placed at the appropriate locations, before being passed to the reproduction system.

Audio objects will be used for dedicated sound events like a ball kicks and a referee whistle blow; a position will be added to other sources such as the TV commentator. However it will not possible to capture and track 45 000 football supporters at once. Therefore the ambience will be recorded as higher order Ambisonics format.

It seems likely that a shift in user expectations may occur when the user becomes an active participant in defining the scene rather than a passive viewer. In the current football broadcast scenario the panning of a camera has no corresponding panning effect on the reproduced audio from the event. In shifting to a viewer-defined scene however the situation is closer to a first-person video gaming scenario where every pan is accompanied by a corresponding shift in the audio scene. Listening tests have been devised and pilot tests carried out within the project to assess this possible paradigm shift in user expectation. A representation of user-controlled FascinatE scene manipulation has been developed giving users control of camera panning within the test shoot panorama. Two scenarios have been presented initially: a static scene with no rotation (the current broadcast norm) and dynamic pan response with both audio objects and rendered ambience rotated according to the user's defined view. Early results from the pilot study, which involved five participants, indicate a likely user preference towards the active participant scenario where the entire sound field, including the audio objects, rotates with the view of the scene. Qualitative evidence from participants in the pilot study suggests that movement of audio objects on the football pitch derived from pitch-side shotgun microphones has a greater subjective effect than rotating the crowd ambience recorded by surround microphones. A full set of tests is planned to determine the optimal audio rendering protocols for the FascinatE system.

6 Conclusion

This paper has outlined the principles of a format-agnostic production system, to support 'virtual re-shooting' of events under the control of either the production team or end users, to suit different devices and user preferences. The concept of a layered scene representation has been introduced, to tailor the resolution of the captured scene to match both the areas of interest and the capabilities of practical production hardware. The first results from an experiment to test these ideas in the context of a football match have been presented.

7 Acknowledgments

The authors would like to thank their colleagues in the FascinatE project for their contributions to the work reported here. They would also like to thank SIS Live for their assistance with the test shoot, and the Premier League for permission to use the football images. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248138.

8 References

[1] SCHÄFER R., KAUFF P., WEISSIG C.: 'Ultra high resolution video production and display as basis of a format agnostic production system'. Proc. IBC 2010, 2010

[2] KROPP H., SPILLE J., BATKE J.M., *ET AL.*: 'Format-agnostic approach for 3D audio'. Submitted to Proc. IBC 2011

[3] NIAMUT O., KOCHALE A., RUIZ HIDALGO J., *ET AL.*: 'Advanced audiovisual rendering, gesture-based interaction and distributed delivery for immersive and interactive media services'. Submitted to Proc. IBC 2011

[4] The ARRI Alexa camera. http://www.arridigital.com/alexa

[5] Faro 'Focus^{3D}' laser scanner. http://www.faro.com/focus/

[6] DAWES R., CHANDARIA J., THOMAS G.A.: 'Image-based Camera Tracking for Athletics'. Proc. IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting (BMSB 2009), May 2009, Available as BBC R&D White Paper http://www.bbc.co.uk/rd/publications/whitepaper181.shtml

[7] GIBB A., THOMAS G.A.: 'Fusion of images using complementary filters and histogram equalisation'. Conf. on Visual Media Production (CVMP2010), November 2010

© The Institution of Engineering and Technology 2011

Selected content from IBC2011

ISSN 2041-5923

Applications of data storage on cinematographic film for long-term preservation of digital productions

C. Voges[1] J. Fröhlich[2]

[1]Technische Universität Braunschweig, Institute for Communications Technology (IfN), Schleinitzstraße 22, 38106 Braunschweig, Germany
[2]CinePostproduction GmbH, Bavariafilmplatz 7, 82031 Grünwald, Germany
E-mail: christoph_vogues@gmx.de

Abstract: Long-term preservation of digital film productions is a challenging task. Conventional storage media with relatively limited lifetimes require data migration in certain time intervals. As a result, today's digital archives need permanent care, maintenance, and thus financial resources. On the other hand, many cinematographic film materials offer an excellent long-term stability, and traditional archiving of the (analogue) film negatives is a both safe and cost-effective long-term storage solution. This paper is about an innovative approach which aims at using these well-established film materials for storage of digital data by means of 'bits on film'. Different setups are suggested to arrange the digital data on the film, optionally also with analog image information on the same medium. A major focus of the paper is on possible applications including a discussion of their practical relevance.

1 Introduction

The large amount of digital data originating from today's film productions is a challenge for digital long-term archiving. In the past decades, film negatives could be used for a reliable preservation of feature films, but digital storage media with a relatively limited lifetime (e.g. [1]) require migration (i.e. re-copying) of the data in certain time intervals. A major disadvantage of this approach is the permanent need of financial resources for such digital archives. On the other hand, many film materials exhibit an excellent long-term stability (e.g. [2]) and are therefore excellently suited for long-term archiving applications. A few years ago, a new interest awakened to use these film materials for storage of digital data (e.g. [3, 4]), also referred to as 'bits on film'. The fundamental idea to store digital data on film is not entirely new (e.g. [5]), and today multichannel sound is digitally stored on film within the Dolby® Digital and the Sony SDDS® systems on cinema film [6]. However, the current 'bits on film' approaches aim at much higher storage capacities and there have already been various contributions to this field of research (e.g. [7–15]). Detailed introductions to this technology are provided in [7, 16, 17]. Owing to its high optical resolution, most of these former approaches are based on microfilm as a storage medium. On the other hand, cinematographic film is a worldwide accepted standard and corresponding reading devices are available and installed all over the world. Recently, it was suggested to use this kind of material as a basis for 'bits on film', especially for digital productions [18]. In this context, a research project is currently being conducted by Technische Universitat Braunschweig, Germany, and CinePostproduction GmbH, Germany.

This paper is about cinematographic film as a medium for 'bits on film' with its specific focus on applications in the field of digital productions. The next Section summarises the aims of the research project and provides an overview of the relevant system parameters. It is followed by a Section that describes the applications and the practical relevance of the project in more detail. The paper ends with a detailed set of conclusions.

© The Institution of Engineering and Technology 2011

2 Project aims and system parameters

The aim of the project is a cost-effective store-and-ignore approach for long-term digital archiving based on cinematographic film. Although basically any kind of data can be stored by using this technology, the main focus is on audio and video data originating from digital film productions. A clear advantage is that both exposure and scanning devices for this type of film are available and installed almost all over the world – originally for digital postproduction purposes. Furthermore, it shall be possible to construct reading devices in the future with considerable effort. Therefore, as a proof of concept, such a reading device is also being constructed within the project based on standard optical components as well as a digital still camera. Depending on the specific film material and the storage conditions, 100 years or more shall be achieved without migration.

For this project, primarily black-and-white film materials (e.g. [19, 20]) have been employed which offer an excellent long-term stability. The ARRILASER [21] has been utilised as an exposure device within the project. Fig. 1 shows a data pattern exposed at a grid space of $d = 10.71$ μm (i.e. the distance between two adjacent data points). To achieve this grid space, the 4K Across Academy exposure format has been used. Since the original grid space of $d = 5.35$ μm of this exposure format would lead to a strong overlap of the data points, only every second pixel was exposed in each direction. As suggested in [7, 14], binary modulation has been employed. Besides the actual data points, a synchronisation pattern can be observed in Fig. 1 that serves to identify the exact position of each data point [9]. File system information has to be added to the digital data to be stored as described in [18]. As a unique property of the medium film, digital as well as analogue data (i.e. photographic images) can be stored on the same medium. Accordingly, it is possible to archive also a human-readable description (e.g. decoding instructions or file format specifications) on the film. Of course, in a practical environment, the film may encounter small damages or dust even if it is handled with care, and forward error correction (FEC) is required to ensure virtually error-free reconstruction of the original data [7, 14, 15]. Therefore, an FEC encoder adds redundancy within the writing process to the digital data that is used by the FEC decoder to correct errors during the read-out process. By neglecting any overhead, e.g. due to synchronisation, file system, and error correction, the so-called gross storage capacity can be calculated. For the above-mentioned exposure parameters this is approximately 193.8 Mbit/m (assuming 1 Mbit = 1024 kbit, 1 kbit = 1024 bit, and 50 frames per metre).

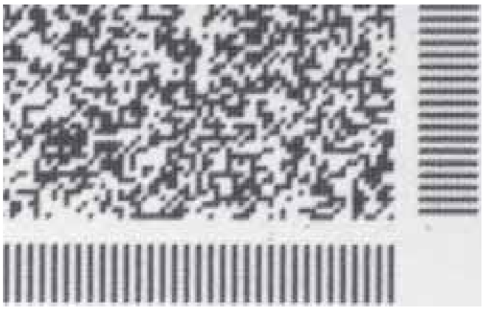

Figure 1 *Microscopic image of a test pattern (binary modulation, d = 10.71 μm) exposed on Fuji Eterna RDS film [20]*

3 Applications and practical relevance

The main mid-term archival media for a feature film throughout the last century was the edited original negative (or duplicate negatives derived from this negative). 'Mid-term' refers to archiving periods from 30–100 years and 'long-term' to 100–500 years in this paper. With the advent of the digital intermediate, today's feature films are archived in both ways, analogue and digital. The analogue archival master is a film typically exposed by means of a laser recorder on colour intermediate film. On the other hand, two different digital formats are employed to store the digital representation: first, the digital source master (DSM) on linear tape open (LTO) tapes and, secondly, a digital cinema package (DCP) [22] which is typically stored in a content management system. When all cinematographic releases will be digital in the future, the colour intermediate film will become obsolete and only digital media will have to be archived. The only real long-term storage solution currently available is the traditional three-strip separation master on black-and-white film material. However, as an example, a feature film of 110 minutes length typically results in six reels for each separation and six additional reels for the sound negative. Altogether, this results in 24 reels for a 110 minute feature film.

In the project that is described in this paper, two entirely new approaches to long-term storage of feature films are proposed by using 'bits on film': The 'hybrid approach' and the 'data only approach' (see Fig. 2).

The 'hybrid approach' is a further development of the traditional separation master process. The centre part of the film is used to store the three separations in an interleaved order. Storing the image information in 2 perf. seems to be an ideal compromise between costs and quality as the resolution of modern black-and-white film stock is higher even compared to original camera negatives with highest resolutions (e.g. modulation transfer curves in [20, 23]). Other current approaches to further developed separation master technology also involve reduction of resolution (e.g. [24]). However, our suggested approach has the clear advantage that soundtrack and images are contained on the

© The Institution of Engineering and Technology 2011

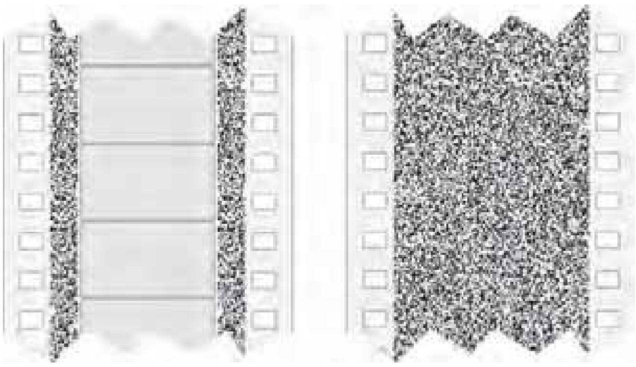

Figure 2 *Schematic illustrations of the 'hybrid approach' (left) and the 'data only approach' (right)*

same negative in the otherwise unused area between the actual picture and the perforation of the film. Thus, the need to create and store a separate set of sound negatives is eliminated. The achieved data rate is sufficient to also include metadata as well as reference frames. Such reference frames (digitally stored and encoded in the X'Y'Z' colour space) can be useful to restore the original colour once the film is scanned back in future. Furthermore, metadata being kept directly on the negative can further help a future restoration if separate metadata databases are lost over the centuries.

The 'data only approach' uses the full super 35 mm frame for data storage. For a feature film in 2K resolution the DSM results in around 2 TByte of digital data versus 100 to 300 GByte for the DCP. The colour space of the DSM may differ among different production facilities and is subject to change owing to technical changes in the workflow of postproduction service providers. This is owing to the currently used 'Cineon' printing density colour space. The ACES/IIF format [25] could solve this issue in future. In contrast to the DSM, the DCP is based on a clearly defined colour space (X'Y'Z') and owing to visually lossless compression the file sizes are only around 5 to 15 percent of the corresponding DSM. Additionally, owing to the wide use of DCP, hard- and software for playback will probably be available for decades. As a result, regarding the current technical situation and economical constraints, the DCP seems to be a very good long-term archival master format. Also, its data rate is very attractive for our suggested approach.

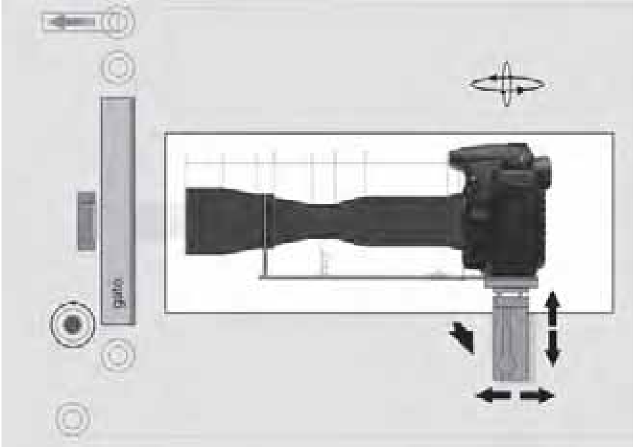

Figure 3 *Schematic illustration of the scanning setup*

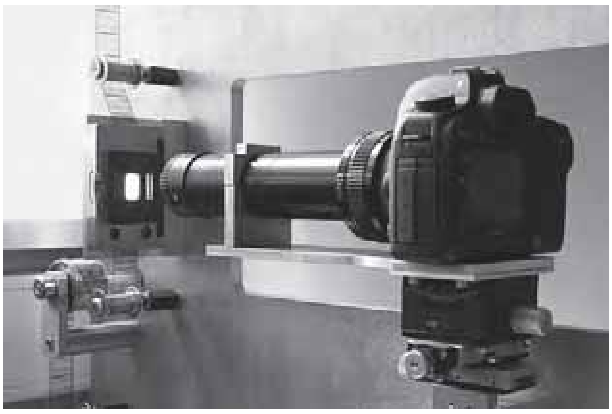

Figure 4 *Camera, macro lens, and film gate of the scanning setup*

When designing an archival system to store data for centuries, it is important to ensure that future engineers are able to recover the digital data without being dependent on today's hard- and software systems. To prove the feasibility of realising a read-out of the data even if film scanners and telecines are no longer available, a scanning device based on standard optical components has been constructed (Figs. 3 and 4). Major components are a Prosumer still photo camera (Canon EOS 5D Mark II), a stepper motor including controller board, an LED unit for constant and uniform illumination, as well as standard computer soft- and hardware. The total costs for the devices add up to only about 5000 Euros.

After the film has been scanned, the content of the film reel has to be interpreted. This can be enabled by printing the decoding instructions as human-readable text on the archived reels. The decoding instructions for the analogue images explain how to transfer photographic densities to luminance and chrominance values. The decoding instructions for the digital part may, e.g. contain the decoder in a written form or even pseudo code to speed up the implementation of decoding systems. The standards for all file formats used and the standards referenced in those standards should also be printed on the decoding instructions part of the film reel.

4 Conclusions

In this paper, 'bits on film' approaches using cinematographic film for long-term storage of digital productions have been discussed. Both exposure and scanning devices for this type of film material are available and installed all over the world.

© The Institution of Engineering and Technology 2011

Furthermore, a film scanning device based on standard optical components has been described. Two approaches have been suggested to store a digital production on the film: the 'hybrid approach' and the 'data only approach', both offering attractive solutions for digital storage of feature films for archiving periods of 100 to 500 years.

The resulting film length of the 'hybrid approach' is only 0.75 times the length of the current mid-term archival master (the recorded colour intermediate and the sound negative) but the black-and-white film has a better long-term stability compared to the colour intermediate film. Compared to a traditional separation master, the 'hybrid approach' requires 0.375 times the film length and storage space. Moreover, the quality of the uncompressed sound is much better compared to the heavily compression used for the traditional digital sound negative. Digital reference frames help to restore the original colour of the film.

The 'all digital approach' using DCPs leads to an amount of 35 mm black-and-white film reels comparable to traditional separation masters. However, the 'all digital approach' does not involve any analog losses. Storing DCPs as they are distributed to cinemas today, will enable future generations in centuries to watch current films in exactly the same image and sound quality as we are watching them today.

6 Acknowledgments

The research project described in this paper is funded by the German Federal Ministry of Economics and Technology (BMWi). The authors would like to thank their colleagues at Technische Universität Braunschweig and CinePostproduction GmbH, especially Gernot Bayer for his contributions to the read-out setup.

5 References

[1] YOUKET M.H., OLSON N.: 'Compact disc service life studies by the library of congress'. Proc. IS&T Archiving Conf., Arlington, VA, USA, May 2007, pp. 99–104

[2] EASTMAN KODAK COMPANY: 'KODAK IMAGELINK HQ, CS, CP and FS microfilms, camera negative microfilm data sheet'. Rochester, NY, USA, 1998

[3] ANGERSBACH C.J., SASSENSCHEID K.: 'Long-term storage of digital data on microfilm'. Proc. IS&T Archiving Conf., Ottawa, Canada, May 2006, pp. 208–209

[4] GUBLER D., ROSENTHALER L., FORNARO P.: 'The obsolescence of migration: long-term storage of digital code on stable optical media'. Proc. IS&T Archiving Conf., Ottawa, Canada, May 2006, pp. 135–139

[5] KUEHLER J.D., KERBY H.R.: 'A photo-digital mass storage system'. Proc. Fall Joint Computer Conf., San Francisco, USA, November 1966, pp. 735–742

[6] HULL J.: 'Surround sound', Ballou G. (Ed.): 'Handbook for sound engineers' (Elsevier Inc., Oxford, UK, 2008, 4th edn.), pp. 1591–1601

[7] VOGES C., FINGSCHEIDT T.: 'Technology and applications of digital data storage on microfilm', *J. Imaging Sci. Technol.*, 2009, **53**, (6), pp. 060 505-1-060 505-8

[8] MULLER F., FORNARO P., ROSENTHALER L., GSCHWIND R.: 'PEVIAR: digital originals', *ACM J. Cultural Herit.*, 2010, **3**, (1), pp. 2:1–2:12

[9] VOGES C., MÄRGNER V., FINGSCHEIDT T.: 'Digital data storage on microfilm – the MILLENIUM project: signal and information processing'. Proc. IS&T Archiving Conf., Arlington, VA, USA, May 2009, pp. 187–191

[10] GIEL D.M., HOFMANN A., SALZMANN W., VOGES C.: 'Digital data storage on microfilm – the MILLENIUM project: hardware realization'. Proc. IS&T Archiving Conf., Arlington, VA, USA, May 2009, pp. 80–81

[11] HOFMANN A., GIEL D.M.: 'DANOK: Long term migration free storage of digital audio data on microfilm'. Proc. IS&T Archiving Conf., Bern, Switzerland, June 2008, pp. 184–187

[12] AMIR A., MÜLLER F., FORNARO P., GSCHWIND R., ROSENTHAL J., ROSENTHALER L.: 'Towards a channel model for microfilm'. Proc. IS&T Archiving Conf., Bern, Switzerland, June 2008, pp. 207–211

[13] VOGES C., FINGSCHEIDT T.: 'A two-dimensional channel model for digital data storage on microfilm', *IEEE Trans. Commun.*, accepted for publication

[14] VOGES C., MÄRGNER V., FINGSCHEIDT T.: 'Digital data storage on microfilm – error correction and storage capacity issues'. Proc. IS&T Archiving Conf., Bern, Switzerland, June 2008, pp. 212–215

[15] PFLUG F., VOGES C., FINGSCHEIDT T.: 'Performance evaluation of iterative channel codes for digital data storage on microfilm'. Proc. IEEE GLOBECOM, Miami, FL, USA, December 2010

[16] VOGES C.: 'An introduction to long-term archiving of digital data on film material'. VDT Int. Convention, Leipzig, Germany, November 2010, to be published

[17] VOGES C.: 'Bits on film – langzeitarchivierung digitaler daten', *FKT (Fernseh- und Kinotechnik), in German*, **65**, (3), pp. 80–84

© The Institution of Engineering and Technology 2011

[18] VOGES C., FRÖHLICH J.: 'Long-term storage of digital data on cinematographic film'. Proc. IS&T Archiving Conf., Salt Lake City, UT, USA, May 2011, pp. 158–161

[19] EASTMAN KODAK COMPANY: 'KODAK panchromatic separation film 2238, technical information data sheet' (Rochester, NY, USA, 1998)

[20] FUJIFILM Corporation: 'FUJIFILM RECORDING FILM for digital separation ETERNA RDS', Datasheet Ref. No. FXX-KB-1006E, 2010

[21] ARRI: 'ARRILASER instruction manual' (Munich, Germany, May 2001)

[22] Digital Cinema Initiative: 'Digital cinema system specification, version 1.2,' Mar. 2008

[23] FUJIFILM Corporation: 'FUJICOLOR NEGATIVE FILM ETERNA Vivid 160', Datasheet Ref. No. KB-0701E, 2007

[24] MCKEE S., PANOV V.: 'Archiving color images to single strip black-and-white 35mm film – the visionary archive process', *SMPTE Motion Imaging J.*, 2011, pp. 24–28

[25] HOUSTON J.: 'Overview and architecture of the image interchange framework'. Presentation at Hollywood Post Alliance Tech Retreat, Rancho, Mirage, CA, USA, February 2008

© The Institution of Engineering and Technology 2011

ISSN 2041-5923

Interview – Anders Prytz

As part of the IBC's focus on young professionals working in the industry we present an interview with Anders Prytz, author of the paper chosen by IBC and the IET as the best young professional contribution for IBC2011.

Tell us a bit about yourself and what you do?

My name is Anders Prytz and I am 26 years old from Drammen, Norway. In June 2010, I graduated in communication technology, specialising in multimedia signal processing at the Norwegian University of Science and Technology (NTNU) in Trondheim. I wrote my Master's thesis in collaboration with Telenor Satellite Broadcasting (TSBc) following two summer internship placements, where I worked with IPTV and video quality, in their broadcasting department. The thesis was entitled 'Video Quality Assessment in Broadcasting' and I have subsequently used this as a basis for my recent article.

I am now employed at TSBc, working in technology within their broadcasting division. TSBc provides extensive television broadcasting services for distribution, contribution and occasional applications to all the Nordic broadcasters and many other broadcasters throughout Europe, using its hybrid network comprised of satellites positioned at 1°West, terrestrial circuits, international teleports and remote earth stations. Additionally, it provides fixed satellite communication and up linking services for data and remote internet applications in Europe and the Middle East.

Our department is responsible for creating and implementing solutions within different areas of broadcasting television services, including DTH, DTT, IPTV and web streaming. We are always investigating and implementing new technologies to ensure we continually provide the best quality service to our customers.

For my part, as a relatively young member of the broadcasting team, I am continually increasing my knowledge both through the accumulated experience of my colleagues, as well as studying and reading academic research and broadcast-related reports. Additionally, we are working with the results presented in my article to work out how we can put the researched material to practical use.

What is your paper about?

TSBc wanted to investigate the relationship between objective measurements of video quality and subjective video quality assessment from a test group. The idea is that video quality measurement could ease the assessment of video quality, with respect to compression settings for distribution. We are currently operating ~400 MPEG −2/−4 encoders of various brands and models. To find the 'perfect' configuration for each brand/type (and software version!) in the respect of video quality is a time-consuming task, using subjective methods. Since video quality is a differentiator in the market, we wanted to investigate if a machine could ease the process of finding the 'perfect' configuration.

We prepared a database of video sequences with different compression settings and showed the sequences to a select test group. A video quality measurement tool was used to measure the video quality of the compressed videos, comparing it to the original. The results from the test group and the measurement tools were used to investigate a possible relationship between the results.

The results proved promising in regards to the direct comparison of man against machine. One of the findings presented in the results, demonstrated strong indications that the objective scores can be related to the objective, once it has attained a certain level. Additionally, another outcome from the results suggested that it is possible to potentially use the initial outcome as guidance when selecting compression settings. However, it should be noted that further research is recommended to achieve a more reliable result.

© The Institution of Engineering and Technology 2011

What brought you into this area and what interests you about it?

After completing the internships at TSBc I was interested in working further with broadcasting. I was very interested in the business and wanted to increase my knowledge within this industry. Additionally, I wanted to my work on the results from the conducted research and focus on the outstanding question relating to how the conclusions could potentially work in practice.

To decide the video quality for 'everyone' is not easy to do. When you have worked with the subject for some time, you are often not representative for the 'common man', regarding choosing what is acceptable video quality and not. When you get used to it you often see errors not noticed by other individuals or groups. The possibility to have a machine telling what the human perception would think of the video quality is an interesting idea. This would also ease our work and give us the opportunity to provide the best video quality to our customers with a reasonable workload. An automated quality assessment tool would also lighten the cost for such a task, since subjective quality assessment sessions are costly and time consuming.

How do you think this area may develop in the future?

A future development in this area would be to be able to use measurement tools to give reliable results for video quality. This could be used to improve television services and ease the work to find the 'perfect' configuration of compression systems. As far as my results concluded, there are still no methods that are good enough to describe what a group of people would choose as the best quality. There are some indications that we, today, can use the methods as a guidance to exclude the worst quality, but choosing the best video quality is still up to people to decide.

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

The main challenge is that it is difficult to make a machine/algorithm that could copy the human perception precisely. The human vision and mind can be too complex and unknown to copy direct. Another problem is that all humans have a complex and unique genetic makeup, so there will always be differences. But to find a solution that could copy the human perception would be the ultimate goal.

The methods I used were only assessing 2D-HDTV content. When you are adding the next dimension, you would have a lot more factors to take in mind, when making algorithms to assess the perceived video quality.

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

This is my first submission to IBC and actually my first ever public submission.

I visited IBC for my first time last year. I was working on TSBc's stand as well as visiting different vendors and contacts we have at work. I was looking around but did not visit the conference session last year, which I am looking forward to attending this year.

Apart from presenting your poster, what else will you be doing at the conference?

During IBC our diaries can become rather full given that we are exhibiting as well as visiting our suppliers and partners. I will try to visit as many sessions as possible. There are many interesting topics at the conference – Connected TVs, 3D and the future of broadcasting in general. Any new ideas, technology and solutions are of interest to develop our own business.

Besides the conference I will work in the exhibition, visiting vendors and contacts, and looking for new technology and solutions to be updated on the technology and what kind of products are there and are in progress.

© The Institution of Engineering and Technology 2011

Selected content from IBC2011. This paper was voted best young professional contribution

ISSN 2041-5923

Video quality assessment in broadcasting – a practical approach

A. Prytz T.Aa. Thoresen

Telenor Satellite Broadcasting AS, Snarøyveien 30, Fornebu, 1331, Norway
E-mail: anders.prytz@telenor.com

Abstract: Broadcasters and service providers are constantly striving to improve the quality of video encoding. To find the best encoder in the market, or the optimal setting of the different encoding parameters, demands a lot of time and resources. The industry has traditionally been skeptical to the use of objective measurements of video quality, owing to the impression of little correlation between the measurements and perceived quality. The test results from a video quality analyser were compared to subjective results from a group of test persons. The objective was to find a correlation between the two methods, so a video quality analyser could be used instead of a test panel, given some restrictions. The authors investigated the objective methods PSNR, DMOS and JND. The results showed poor correlation between PSNR and subjective quality scores. DMOS and JND gave better results. However, content with high complexity (spatial and temporal) gave poor correlation.

1 Introduction

Telenor Satellite Broadcasting (Telenor SBc) transmits over 250 digital TV channels and 70 radio channels across the Nordic countries and throughout Europe by satellite. Telenor SBc also encodes numeral services for IPTV and terrestrial transmissions in Norway.

The company has numerous encoders from different vendors and with different versions. To improve the picture quality for a given encoder, at a specific bitrate, is a task that consumes a lot of time and effort. Objective measurements that can be related to subjective perception of the quality could potentially reduce the cost related to quality testing.

Today, no objective methods have been proven accurate enough to be used as a measurement related to human perception. Video Quality Experts Group (VQEG) has done research that gave no recommended method/ measurement, see Webster *et al.* [1] and Corriveau *et al.* [2].

In this paper we evaluate full-reference objective methods that could assist or completely perform the quality assessment of TV channels. Objective measurements were conducted by a video quality analyser (VQA) system, which had three methods implemented.

The relationship between the subjective and objective quality measurements was investigated, with the help of subjective evaluation sessions which was analysed and compared with the objective measurements.

2 Methods

2.1 Video test content

To cope with a broadcasting scenario, we wanted to have different kinds of video content to reflect possible content shown on a TV channel. The video content was 10 seconds long and originated from 'The SVT high definition multi format test set' [3] and from a football match.

Four videos were captured from [3] and two scenes were captured from the football match. 'The SVT high definition multi format test set' is a well-known test sequences in the industry and the football was recorded on HDCAM SR directly in the OB-van, to represent the sports part of television. The content is described in more detail in the Master's thesis 'Video Quality Assessment in Broadcasting' by Prytz [4].

© The Institution of Engineering and Technology 2011

All the content was native 1920 × 1080i at 25 Hz to represent the usual HD resolution. All sequences were exactly 10 seconds long, as recommended in ITU [5].

The sequences provide a mix between very complex scenes and more 'normal' television content.

The encoding was H.264/AVC (MPEG-4 Layer 10). A number of settings were varied during the tests; bitrate, horizontal resolution and GOP structure. The settings can be seen in Table 1.

2.2 Subjective methods

The subjective evaluation was performed in accordance with the recommendations from ITU-R BT.500 [5]. The room used to conduct subjective quality evaluation was modified to cope with the ITU recommendation [5] for a home environment.

The subjects who participated in the quality evaluation were without any training or experience in picture evaluation. The majority were students and employees at Norwegian University of Science and Technology (NTNU). The average age was 25.7 years with 30 people participating.

The evaluation session consisted of an instruction, training session, pause/questions, and assessment session. The video sequences were of ten seconds, and was shown continuously, single stimulus, with a three second grey screen in between. The total assessment time did not exceed 20 minutes. There was no comparison in the assessment, and voting was done on a 1–5 scale. A hidden reference video sequence (uncompressed) was included in the voting process.

Table 1 Different compression settings used to encode the test video content

Code	Bitrate	GOP length	GOP Structure	Hor. Res.
A	10	12	IP	960
B	15	12	IP	960
C	10	12	IP	1920
D	15	12	IP	1920
E	8	12	IBBBP	960
F	10	12	IBBBP	960
G	15	12	IBBBP	960
H	8	12	IBBBP	1440
I	10	12	IBBBP	1440
J	15	12	IBBBP	1440
K	8	12	IBBBP	1920
L	10	12	IBBBP	1920
M	15	12	IBBBP	1920

2.3 Objective methods

Three different objective measurement methods were tested:

- Peak signal-to-noise ratio (PSNR). PSNR is a quality assessment method and a direct measurement on how much the processed image/video is different from the original, pixel by pixel. PSNR is given in decibel and is a logarithmic function calculated by the possible number of colour representation and the mean squared error.

- Just-noticeable differences (JND). The Sarnoff JND visual model [6], is a method developed by Sarnoff/Tektronix, which predicts the subjective rating for a group of human testers. The method analyses the image for macroblocks, blur, luminous variations etc. It predicts a score that is correlated to the JND scale using the VQEG database.

- Different mean opinion score (DMOS). DMOS is the name of one of the objective assessment method on the VQA system that gives results 'directly' comparable to subjective results. The measurement is based on MS-SSIM, see Wang *et al.* [7] and VideoClarityInc [8], which gives a score that, are mapped to DMOS through a polynomial fitting function.

3 Results and discussions

3.1 Subjective tests

The most definite result from the subjective quality assessment sessions was that increased bitrate gives increased picture quality. From Fig. 1 we see that the increasing value for MOS for each colour is directly linked to bitrate.

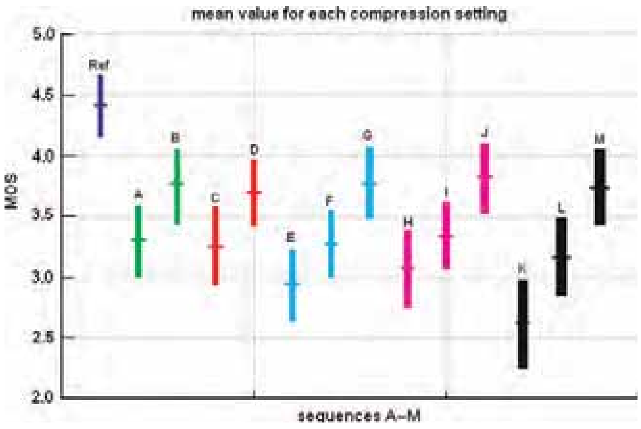

Figure 1 *Different video compression setting with subjective result (for all test video contents)*

Within each colour, the only variable is the bitrate

© The Institution of Engineering and Technology 2011

The mean result for all votes was 3.44, which is a little higher than expected. A golden rule is to achieve a total mean at 3 when using a 1–5 scale.

The series E, H and K varies only in horizontal resolution. It is interesting to note that K has significantly lower MOS score compared to E and H. This observation indicates that video quality can be improved by reducing the horizontal resolution if the available bitrate is limited. The same can be seen in the series F, I, L and G, J, M, but less significant owing to the high bitrate used in these series.

3.2 Subjective against objective results – without constrains

The direct correlation between subjective and objective results was calculated with linear and fitting functions (polynomial and exponential). The polynomial function achieved the best results, and is also the function used by VQEG in the latest test plan [9]. The relationship is given with a number ≤1, where 1 describes a distinct relationship.

PSNR achieved a correlation of 0.37 at its best, which was the lowest in total. This result can be seen in Fig. 2. While PSNR is the most used and most well-known objective measurement, it gave not a distinct relationship to subjective results.

The next best method was the JND method, which achieved a correlation of 0.55.

DMOS showed the best relationship between objective and subjective result, with a correlation of 0.69.

A correlation of 0.69 is still not good enough to give a distinct relationship, but DMOS show some promising results, taken into account that PSNR was as low as 0.37. An attempt to improve the relationship between objective and subjective results was made by using the spatial and temporal information of the video content and is explained below.

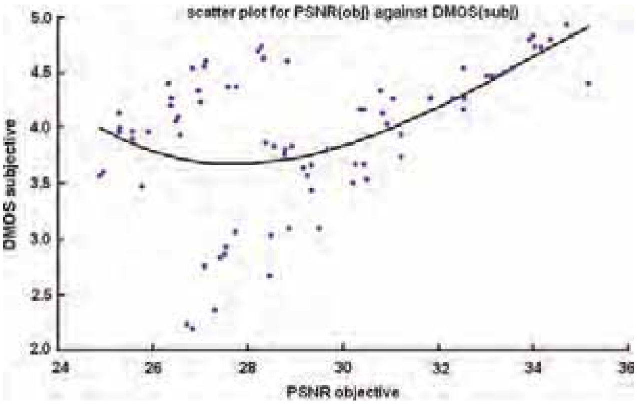

Figure 2 *Objective PSNR against subjective DMOS for all test video contents*

Correlation is 0.37

3.3 Spatial and temporal information

To find a differentiator between the different contents, we tried to use the spatial information (SI) and temporal information (TI) of the video test content. In short, spatial information is a measurement of similarities and differences across a frame and temporal information the same between frames. This is a calculation of the complexity in each video sequence performed by the VQA system and the measurement method is in compliance with ITU-R P.910, ITU [10].

3.4 Subjective against objective results – with constrains

As a result of the lack of a distinct relationship between subjective and objective results, we tried to split the sequences into groups defined by the complexity of the spatial and temporal information. This was done in an attempt to improve the results we achieved when all the sequences in the test was analysed as one group.

The reference video content (uncompressed) was used to calculate the spatial and temporal information complexity. The calculations resulted in Fig. 3, which have the mean values indicated for each axis (indicated by the line crossing the axis). The classification of each content shows that the content have high variation in both the spatial and the temporal domain.

For PSNR, the content was sorted into two groups, divided by the mean value of spatial information. The result of the separation was that the low SI group achieved a 0.93 correlation, while the high SI group got 0.85. This result was a great improvement over the aggregated correlations and indicates that video content classification might be useful.

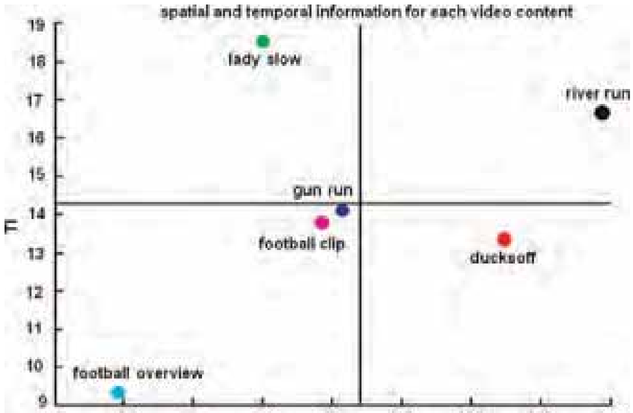

Figure 3 *Spatial and temporal information for the different test video contents*

© The Institution of Engineering and Technology 2011

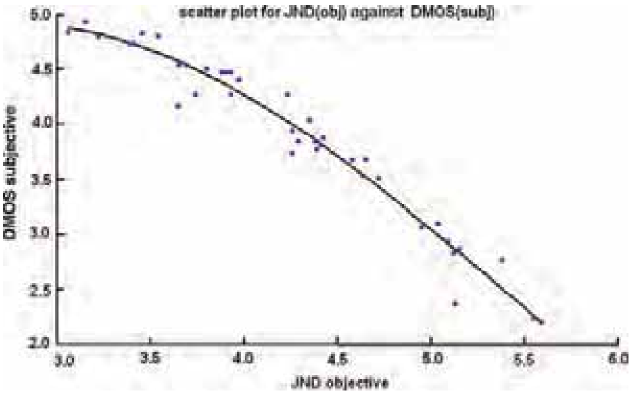

Figure 4 *Relationship between subjective DMOS and objective JND with a polynomial function for the low SI low TI group*

Correlation at 0.98

The content was separated into four groups for JND. The groups are the four squares, seen in Fig. 3, divided by mean TI and SI. The low SI – low & high TI groups got a 0.98 correlation. However the high SI groups gave a correlation on 0.91 and 0.82 for low and high TI, respectively. Fig. 4 shows the correlations between the subjective and the objective results with plots and the given correlated polynomial function.

The content was also separated into four groups for DMOS. These groups were created based on the spatial information, <60, 60–85, 85–105, >105. While JND showed a significant 0.98 correlation, DMOS show a lower, but more stabile correlation. With the results of 0.94, 0.96, 0.91 and 0.92 for the respective groups, DMOS is the most stabile objective method with all groups above 0.90.

4 Conclusions

One very obvious conclusion is the fact that the single most important factor in picture quality of encoded video content is the bitrate.

Reducing the horizontal resolution can improve the perceived video quality significantly if the available bitrate is limited.

As a total, the DMOS method showed the best overall result when each SI/TI group was evaluated separately. JND showed the best result, with a correlation at 0.98, but failed to achieve this result for all its SI/TI groups. PSNR gave too poor results to be evaluated as a candidate for reliable measurements.

A relationship between subjective and objective results might be dependent on the content information complexity. There is a great improvement in the correlation when the content is grouped by their information complexity, compared to the correlation for all contents as one group.

We have not been able to find a good correlation between the objective and subjective measurement with content with both high spatial and temporal information. We are not able to evaluate if this is owing to underperformance in the subjective measurement methods, or the test subjects' inability to assess picture of poor quality, but we can conclude that too much spatial and temporal information in the video content might disturb a quality assessment.

The result of the research in this paper gives an indication that the relationship between objective and subjective quality assessment might be affected by the spatial and temporal information complexity of the video content.

Regarding the lower correlation for content with high information complexity, compared to low information complexity, some questions are still unanswered:

- Is the selected content to complex and the encoded result too poor for people to assess the video quality?

- Or does the objective methods fail to assess this material?

5 Acknowledgments

The authors would like to thank TV2 Norway for the use of their content and everybody who participated in the subjective evaluation.

6 References

[1] WEBSTER A., SPERANZA F.: 'Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I'. Technical report, VQEG, September 2008

[2] CORRIVEAU P., WEBSTER A.: 'Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II'. Technical report, VQEG, August 2003

[3] HAGLUND L.: 'The SVT high definition multi format test set'. Technical report, SVT, February 2006

[4] PRYTZ A.: 'Video quality assessment in broadcasting'. Master's thesis, NTNU, 2010

[5] ITU: 'ITU-R BT.500-11: Methodology for the subjective assessment of the quality of television pictures,' Question ITU, 2002, vol. 211, p. 11

[6] LUBIN J., FIBUSH D.: 'Contribution to the IEEE standards subcommittee: Sarnoff JND vision model', 1997

© The Institution of Engineering and Technology 2011

[7] WANG Z., SIMONCELLI E., BOVIK A.: 'Multiscale structural similarity for image quality assessment'. Conf. Record of the 37th Asilomar Conf. on Signals, Systems and Computers, 2003, vol. 2, 2003

[8] VideoClarityInc: 'Clearview users' guide', webpage, http://www.videoclarity.com/PDF/ClearViewUser Guide.pdf, April 2010

[9] CERMAK G., THORPE L., PINSON M.: 'Test plan for evaluation of video quality models for use with high definition tv content'. Video QualityExperts Group (VQEG), 2009

[10] ITU: 'ITU-T Recommendation P.910 subjective video quality assessment methods for multimedia applications', September 2009

© The Institution of Engineering and Technology 2011

ISSN 2041-5923

Introduction to *Electronics Letters*

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

Each year the executive committee of the IET community for Multimedia Communications comes together to select a small number of papers from the relevant content of *Electronics Letters* to appear in this publication. This year three Letters have been selected, of which the first two deal with the processing of video media. The first of these looks at an issue created by the growing amount of 3D content – the retrieval of relevant stereoscopic content taking into account not just traditional 2D image features, but also the depth information available in 3D media. The second Letter deals with using video content across a range of very different devices, specifically different screen sizes – how do you take soccer video shot for viewing on the traditional main family-size home screen and optimise it for viewing on the much smaller area of a smartphone screen, automatically? The final Letter deals with human–computer interaction through gesture recognition, which current trends suggest is emerging as an important area both in home entertainment and (as suggested by the IBC paper of Jung *et al.* on page 11) interactive advertising.

Electronics Letters is now in the second year since a relaunch, which saw the inclusion of a magazine-style news section (freely accessible through our homepage) including feature articles based on Letters in the issue and author interviews. As a taste of this content each of the above Letters is preceded by an author interview providing more background and insight on the work reported in their *Electronics Letters* paper.

We hope that these Letters will provide you with an example of the kind of broadcasting industry relevant, new and interesting research published in *Electronics Letters*.

The *Electronics Letters* editorial team.

[1] www.theiet.org/eletters

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

© The Institution of Engineering and Technology 2011

Interview – Yue Feng

What is your current role and how did you get there?

After graduating from the University of Shanghai for Science and Technology (USST) with a BEng in Computer Science in 2003, I moved to the UK and obtained a PhD in Electronic Imaging and Media Communication from the University of Bradford. My PhD research focused on image and video processing, developing methodologies for converting 2D media to 3D, and intelligent video management and retrieval. In 2007, I joined the Multimedia Information Retrieval group at the University of Glasgow as a postdoc research associate in retrieval. In 2010, I moved back to Bradford University, where I continued working on stereo visuals retrieval and 3D media processing. I now work at King's College London as a postdoc researcher in magnetic resonance imaging (MRI) and stereo image processing for medical images. Along the way I started working with Dr. Jinchang Ren, currently with the Centre for excellence in Signal and Image Processing (CeSIP) at the University of Strathclyde, Glasgow. This work is the latest of our collaborations.

What was the motivation for this work?

Thanks to the stereo visual content and device producers, stereo media programmes including 3DTV and games have received increasing attention in the last decade, as a way to expand user experiences. Given the dramatic increase in the commercial interest in, and popularity of, stereo programmes, especially since the 3D film 'Avatar' was released, more and more stereo programmes are produced everyday. Inevitably this drives a requirement for efficient retrieval of such content cope with user demand and improve efficiency in video production.

What is the main challenge in doing this kind of work for stereo visuals as opposed to single view?

The fundamental difference is that stereo video is taken by a binocular camera and so uses a pair of images to represent a scene, where 2D visual is taken by a monocular camera and only has a single image to represent. The extra channel introduces more depth cues. The most challenging part of the work is how to find the best proper depth cues and integrate these with traditional 2D content-based image features for retrieval. The depth and the content-based image features are from two different feature spaces, so smooth and organic integration of these two is critical in such a system.

What has been achieved in the work reported in your *Electronics Letters* paper?

Our Letter reports a framework for content-based stereo image/video retrieval, offering a general solution for image retrieval in stereo content. It combines traditional visual features with depth features to retrieve relevant results. The 2D image features and the depth are used to determine visual similarity and depth consistency between the query and the candidate results respectively. A re-ranking scheme is then introduced to combine the two-similarity estimations together to improve the retrieval accuracy.

What applications do you have in mind?

By addressing the problem of stereo media retrieval, I think this could be widely applied in media content management systems, where searches for similar video clips or shots are needed. In addition, it could also be applied in search engines for stereo content retrieval. The search/retrieval engine, which is now widely used in daily life, has changed the world dramatically. Therefore, I believe stereo media retrieval applications will have a huge impact on the future media world, especially, if 3D media continues to become more dominant.

What have you been doing to develop the work since your Letter?

Currently, Dr. Ren and I are working on visual place classification using stereo video sequence as input. It aims to classify video sequences into different scenes according to the environment. We believe different environments will have differences in visual or depth similarity.

What else are you working on at the moment?

Nowadays, researchers cannot focus on one area. Apart from the stereo content research, I am also working on MRI research in autism, where the aim is to build up a brain model using MRI image series to extract differences in the brains of people with and without autism. The results will then be used for doctors to diagnose the disease. Dr. Ren's interests cover video content analysis, 3D computer vision, visual surveillance, archive restoration, medical imaging and machine learning, but he now focuses mainly on hyperspectral imaging.

How do you think this field will develop? What would you like to see?

There will be a bright future for this field because I believe many companies have seen the commercial potential of 3D. More research has been initiated for developing new hardware, software and 3D content.

In addition, online and offline search services will keep developing to allow people to search information more effectively. Therefore, I think 3D content searching will be one of the next jobs in their agenda. In the future, I would like to see a retrieval engine, which utilises features from different domains. For instance, given a multimedia query like a video clip from a 3D movie, the engine would acquire the audio, visual, and depth features from the query example to perform a search in the database. I also think online applications over the internet via smartphones will develop rapidly and complement conventional applications in the TV/movie and media industry.

© The Institution of Engineering and Technology 2011

Selected content from *Electronics Letters*

ISSN 2041-5923

Generic framework for content-based stereo image/video retrieval

Y. Feng[1] *J. Ren*[2] *J. Jiang*[1]

[1]School of Informatics, University of Bradford, Bradford, UK
[2]Centre for Excellence in Signal and Image Processing, University of Strathdyde, Glasgow G1 1XW, UK
E-mail: yfeng2@brad.ac.uk

Abstract: With the increasing number of stereo images/videos in commercial markets, the demand for content-based image retrieval (CBIR) to deal with stereo media becomes urgent. To meet this demand, a novel framework is proposed where depth cues are extracted from stereo pairs and employed in a re-ranking scheme to refine results from conventional CBIR. Experiments show the proposed method yields promising results in retrieving stereo content.

1 Introduction

Nowadays, stereo images/videos are more and more popular as they enable immersive experiencing in a wide range of applications, such as 3DTV/movies for entertainment and stereo geographical information systems, etc. [1]. However, how to effectively retrieve content of interest from such stereo media has not been investigated. Since a stereo pair rather than a single image is utilised to represent a scene, this forms the basic difference between stereo image retrieval and the conventional content-based image retrieval (CBIR) system, i.e. how to extract and utilise valuable features to represent stereo images for their effective search and retrieval. Although a number of CBIR systems have been proposed to retrieve images/videos [2], using visual features like edge, colour, and homogenised texture, few of them are specifically designed for stereo material regarding its nature of rich in-depth information. Tailoring the architecture specifically for stereo image/video retrieval, we adopt a re-ranking model where visual features provide primary evidence for retrieving relevant documents, while the disparity features extracted from the stereo pairs offer complementary clues to refine the results further. Consequently, our work is the first attempt to solve this problem. Since the proposed framework can be used to extend existing CBIR systems, intrinsically there is no difficulty in terms of consistency and portability to constrain its wide application.

2 Similarity measure in CBIR

Based on evidence from the best-performing video retrieval systems in TRECVid 2007 and 2008 search and retrieval tasks [3], MPEG-7 standard low-level visual cues are the top performing and most popular features for retrieving relevant documents in conventional CBIR, where text and audio features are absent. In this framework, edge histogram and colour structure descriptors are employed.

Based on a suitable distance measure, the similarity is determined by means of comparing the feature vector $\boldsymbol{x}_q = (x_{q1}, x_{q2}, \ldots, x_{qK})$ of a query image with the feature vectors $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jK})$ of the *j*th image in the database. Thus, a combined similarity is obtained as $s_{qj} = g(f_1(x_{q1}, x_{j1}), f_2(x_{q2}, x_{j2}), \ldots, f_i(x_{qK}, x_{jK}))$ with respect to individual features, where $f_i(x_{qi}, x_{ji})$ represents the similarity for the *i*th feature. Function *g*() is to fuse separate similarities for a combined one, where weighted sum is a straightforward form used for this purpose.

3 Re-ranking using estimated disparity maps

Using the extracted similarity measurements above, conventional CBIR retrieves a list of candidate images, where the most relevant element is the one with the highest similarity to the query image. By analysing the obtained

© The Institution of Engineering and Technology 2011

retrieval results, it is found that much false matching happens owing to inconsistent scene depth. For instance, a meeting room may match with an office scene sharing similar appearance in terms of decoration and furniture layout, although their scene depths are significantly different. This shows how essential depth information can be used to complement visual features in the effective retrieval of images and videos, especially for stereo ones.

In stereo vision applications, it is well known that depth is inversely proportional to disparity provided that cameras are parallel and epipolar lines are horizontal [4]. Therefore, the problem of finding depth cues is equal to extraction of a disparity map. Fig. 1 shows an example of one left-image and its associated disparity map, where brighter contents are closer to the camera and vice versa. Actually, the disparity map is applied as a depth cue to re-rank the results from conventional CBIR as follows. Let D_q and D_j be the estimated disparity map for the query image and the jth image in the database. The consistency of D_q and D_j, Φ, are each then used to adjust the obtained visual similarity as $s'_{qj} = \Phi(D_q, D_j)s_{qj}$. For efficiency, re-ranking is applied only to the first 10 to 20% of candidates from conventional CBIR, and the retrieved results are then re-ordered according to adjusted similarity measures s'_{qj}.

In our implementation, histograms of D_q and D_j are determined, and this enables us to employ histogram similarity to measure the consistency of the two depth maps, where diffusion distance in [5] is used as Φ to measure such consistency. There are two reasons to apply diffusion distance here: one is that the disparity map is similar to a temperature field in a diffusion process; the other is that diffusion distance is robust to deformation, lighting changes and noise [5]. In fact, much improved results have resulted from using this distance measurement, and results are presented in the following Section.

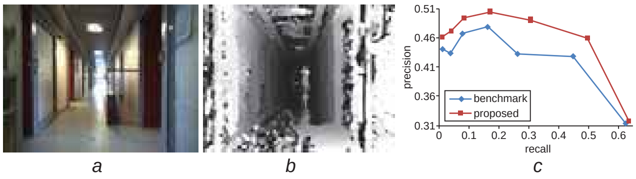

Figure 1 *Left-image of stereo image pair (Fig. 1a), its disparity map (Fig. 1b), and precision-recall curves of retrieved results (Fig. 1c)*

4 Experiment results

In our experiments, stereo video frames in the robot vision retrieval task of ImageCLEF 2010 [6] are utilised to validate the effectiveness of the proposed approach. The task is to search relevant images in a training set to decide the location of a mobile robot, where stereo query frames are real-time captured when the robot with mounted cameras moves in the scenarios. Ground truth like robot locations among several venues (including corridor, kitchen, small office, large office, bathroom, printer area, recycle area as well as meeting room and library) is provided for quantitative evaluations.

Basically, two groups of results were obtained and compared in our experiments to evaluate the proposed approach. One is retrieval results using only visual features, i.e. the same way conventional CBIR works. The other is refined results from the first group with proposed disparity-based re-ranking. It is worth noting that all the results are generated by us to validate the proposed method. For each group of results, two measurements, recall and precision, are computed and are illustrated in Table 1, where precision and recall rates from the first m results are attained for comparison. As seen, thanks to the proposed re-ranking scheme, improved recall and precision rates are achieved for all possible values of m. In addition, recall-precision curves are plotted in Fig. 1 to compare further the performance of our proposed approach with the benchmark system. As seen, using disparity/depth for re-ranking indeed has significantly improved the retrieval performance.

Table 1 Comparison of retrieval results from benchmark (no re-ranking) and our system (with re-ranking) using robot vision data in ImageCLEF 2010

m	Benchmark, %		Proposed, %		Improvements, %	
	Recall	Precision	Recall	Precision	Recall	Precision
At 10	0.98	44.13	1.01	46.25	0.03	2.12
At 40	3.64	43.44	3.93	47.22	0.29	3.78
At 80	7.74	46.80	8.17	49.47	0.43	2.67
At 160	16.14	47.90	17.09	50.51	0.95	2.61
At 320	26.24	43.29	30.68	49.11	4.44	5.82
At 640	44.72	42.83	49.61	45.94	4.89	3.11
At 1280	62.40	31.31	63.68	31.85	1.28	0.54

© The Institution of Engineering and Technology 2011

5 Conclusion

A novel framework for content-based stereo image/video retrieval is proposed, where the constraints of consistent disparity/depth are employed in a re-ranking scheme to refine results from conventional CBIR. Using histogram-like diffusion distance to measure such consistency, significant improvements are achieved in terms of better recall and precision rates. Since the proposed framework can be utilised as an extension to enable existing CBIR systems to deal with stereo contents, it has great potential to be applied in the coming boom of 3DTV/movies.

6 References

[1] KRISHNAPURAM R., MEDASANI S., JUNG S.-H., *ET AL.*: 'Content-based image retrieval based on a fuzzy approach', *IEEE Trans. Knowl. Data Eng.*, 2004, **16**, (10), pp. 1185–1199

[2] FENG Y., URRUTY T., JOSE J.: 'A novel retrieval framework using classification, feature selection and indexing structure', *Lect. Notes Comput. Sci.*, 2010, **5916**, pp. 731–736

[3] TRECVID 2008 – Goals, Tasks, Data, Evaluation Mechanisms and Metrics, http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2008

[4] BARNARD S., WILLIAM T.B.: 'Disparity analysis of images', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1980, **2**, (4), pp. 333–340

[5] LING H., OKADA K.: 'Diffusion distance for histogram comparison'. Proc. IEEE Conf. CVPR, New York, NY, USA, 2006, vol. 1, pp. 246–253

[6] http://www.imageclef.org/2010/robot

© The Institution of Engineering and Technology 2011

Interview – Li Gao

What is your background and what is your current role?

I received my BSc and MSc in Electronic Engineering from Beijing Institute of Technology, in 2001 and 2004, respectively, and my PhD in Signal Processing from the Institute of Acoustics, Chinese Academy of Sciences (CAS), in 2007. From 2007 to 2009, I was an assistant researcher in the institute and since 2009, I have been an associate researcher here. My research covers image/video processing, pattern recognition and machine learning, and signal processing.

What attracted you to working in this area?

I find the question of how to improve the user's perception of content through a small display an interesting problem. Video retargeting or re-framing is one of the key technologies and video object segmentation is an important basis of this. I had done research in video object recognition, shot segmentation, and coding. Recently new research problems in these subjects have emerged with the increasing popularity of video browsing through mobile devices. So, based on my previous work, my interests began to move into this area.

Why is this kind of re-framing of video needed?

When presenting high-resolution videos designed for large screens on small mobile devices, direct down sampling for mobile display capacity may lead to over shrinking of image objects. This makes the objects of interest of unacceptable quality or even unrecognisable to the user. Also, in the long duration, long distance shots of broadcast soccer video, tiny objects, such as a soccer ball and players, may not be comfortably perceptible on a small panel. Re-framing is necessary to improve the visual perceptions of users. Soccer video is distortion-intolerant since the changing of spatial relationships among the ball and players caused by non-homogeneous re-framing leads to incorrect understanding of events, so we use cropping based re-framing.

What has been achieved in the work reported in your *Electronics Letters* paper?

Methods have been proposed to re-frame soccer videos before. The simplest methods detect the ball and make it obvious through colour and size. Another, proposed in 2007, used ball location and speed to determine region of interest (ROI). However, a fast moving ball is easily missed in ROI. Moreover, players and other elements like goal, field line, referees, that haven't been considered are also important to understanding soccer events.

Thus, semantic events may not be well exhibited in ROI when only considering the ball. We propose a content-aware cropping-based retargeting for soccer long-shot frames, where a so-called fuzzy visual perception is introduced to retain as much content of interest as possible in the ROI. Our method has three unique features. First, we introduce visual perception constrained by four rules and simulated by a fuzzy inference system (FIS) for the retargeting. The FIS collects human knowledge for decision making and fuses visual perception rules for ROI detection. Secondly, to achieve content-aware retargeting, information about both the ball and players is considered for ROI detection. Thirdly, we use a fuzzy set model to derive a visual perception model for ROI. Through visual attention features, visual perception constraints are quantised in an easy to understand way. In addition, we attempt to quantify each object's contribution to the ROI by calculating visual attention value.

However, in terms of both object detection and the domain-specific knowledge-based ROI detection, the computation complexity is not high. So its use would not require new hardware or capabilities in current devices.

What differences are there in applying this to other types of video footage?

This is a top-down approach to 'bridge' the gap between low level features and high level semantics. The reason for selecting sports video to research is that it has a rich events structure. Soccer video comes with more prior knowledge than non-sports sequences. Using domain-specific knowledge makes our top-down approach realisable. Through prior knowledge of ground colour, field line, and ball etc. many objects such as ball, players, goal, and field lines can be extracted automatically. The high level ROI detection also uses domain-specific knowledge. To this end, we did a detailed, large-scale user study to get statistical domain-specific knowledge about user preferences on soccer video content, which included the interesting content in each event. Then we used perception clues and FIS to derive a model that quantifies interesting content and, further, each object's contribution to the ROI. But in general non-sports videos, the prior knowledge about objects and preference of users is difficult to obtain or define in advance.

What are the next steps in developing the work?

For soccer video, other elements and perception clues, such as goal, field lines, penalty point, corner flag and events will be modelled and added into the FIS. Right now I am researching events-based ROI detection (such as 'goal', 'dribbling', 'free kick', 'corner ball', 'flank pass', 'penalty kick' event). This will make the problem more detailed and more in accordance with viewers' perceptions of soccer videos, allowing the model to adjust adaptively according to different events. Extending the work beyond soccer could be realised in two ways. First, through simple feedback to get each user's focus in viewing video, prior knowledge about user preferences could be obtained. Secondly, videos may be classified into different types or different scenes. Through survey and learning in each case, we could get prior knowledge about user preferences in each type or scene. This knowledge could then be used to derive a visual attention model using a similar approach to this work.

© The Institution of Engineering and Technology 2011

Selected content from *Electronics Letters*

ISSN 2041-5923

Content-aware broadcast soccer video retargeting using fuzzy logic

L. Gao[1] *M. Xu*[2] *S.F. Yan*[1] *M.G. Liu*[1] *C.H. Hou*[1] *D.H. Wang*[1]

[1] *Institute of Acoustics, Chinese Academy of Sciences, 21 Beisihuanxi Road, Haidian, Beijing 100190, People's Republic of China*
[2] *School of Computing and Communications, University of Technology, Sydney, Australia*
E-mail: future_gao@hotmail.com

Abstract: A content-aware video retargeting method is proposed for playing broadcast soccer video in small displays. Four visual perception clues are predefined based on soccer game-specific knowledge and modelled by visual attention features firstly. Then, a fuzzy logic inference system is proposed to estimate visual attention values (AVs) of ball and players by fusing attention features. AVs are later used to determine the region of interest (ROI) of each frame. Finally, a retargeted video is generated by the ROI of each frame with polynomial curve fitting for temporal smoothing. Both subjective and objective evaluation results are promising.

1 Introduction

An original video stream for TV or HDTV needs to be transformed into thumbnail videos for playing on small displays, such as a mobile phone. Early methods of direct down-sampling may bring viewers an uncomfortable watching experience [1], especially when small objects appear in original videos. Recent methods of video retargeting can be summarised into cropping based resizing [1, 2] and non-homogeneous resizing [3, 4]. Cropping based methods only remain in the region of interest (ROI). Non-homogeneous resizing brings distortion problems by allocating high resolution to important objects compared to non-important regions. Sports video is distortion-intolerant since the changing of spatial relationship among ball and players due to non-homogeneous resizing leads to a wrong understanding of sports events. It was also mentioned in [3, 4]. As one of the most popular broadcast sports, soccer video is used in this Letter. In long-view shots, objects (soccer ball and players) are too small to be recognised if the whole frame is shown on a small display. Therefore, we focus on soccer long-view shots and choose the cropping based method for retargeting. Normally, an ROI is selected according to a saliency map [3] to generate a resized video without considering video content. In [1], soccer ball and ball moving speed were used to determine the ROI. However, we find that a fast moving soccer ball is easily missed in the ROI by using the method in [1]. Besides soccer ball, players who have not been considered in [1], are also important clues to interesting content. In this Letter, we propose a content-aware cropping based retargeting for soccer long-view shots. The ROI is determined by four perception clues related to soccer ball and players, which are predefined based on soccer semantic constraints and domain-specific knowledge. The perception clues can be modified based on users' preference. Rule fusion is also a concern in this Letter. Three unique features are summarised as follows: (i) compared to [1], to achieve a content-aware retargeting, both soccer ball and players contribute to ROI detection; (ii) visual perception is modelled by visual attention features; (iii) fuzzy logic which collects human knowledge for decision making is introduced to fuse inference rules for ROI detection. The proposed method can be easily extended to other sports domains by applying domain-specific perception clues.

2 Overview of our retargeting method

Our goal is to find the optimal cropping region, i.e. the ROI, to maximise interesting contents within the limited cropping region. Visual attention features (VAFs) are first extracted to model four predefined visual perception clues. Secondly, fuzzy logic is applied to calculate the attention value (AV) of each object (i.e. ball or player) by fusing VAFs according

© The Institution of Engineering and Technology 2011

to fuzzy rules. Finally, based on the AV of each object, content-aware ROI is determined by including maximal objects with high AVs. Retargeting video is further generated by the ROI of each frame.

3 Visual perception clues and attention features extraction

Ball and players are detected using our previous method [5]. Then, VAFs are extracted according to visual perception clues individually. In this Letter, four visual perception clues are designed (but without limitation) based on a wide study on users' preference. 1. The soccer ball is the most important, which should be displayed in the ROI. 2. Players close to the soccer ball are important since they might be able to control the ball. We use the distance from each object to the ball as the VAF to describe this clue: $\mathbf{DB} = [DB_{ball}, DB_1, \ldots, DB_k, \ldots, DB_M]^{\mathrm{T}}$, where DB_{ball} is 0. $DB_k = ((x_k - i_b)^2 + (y_k - j_b)^2)^{1/2}$, where $(i_b, j_b)^{\mathrm{T}}$ is the ball's position, and $(x_k, y_k)^{\mathrm{T}}$ is the position of the kth player. 3. The ROI should contain as many as possible players because players' formation and activities (such as pass, offside and shoot) are very important to convey soccer events. The distance from each object to density centroid, $\mathbf{DC} = [DC_{ball}, DC_1, DC_2, \ldots, DC_k, \ldots, DC_M]^{\mathrm{T}}$, is used to measure the above clue. $DC_k = ((x_k - i_c)^2 + (y_k - j_c)^2)^{1/2}$, and $(i_c, j_c)^{\mathrm{T}}$, is the density centroid of ball and players, where $i_c = \left(i_b + \sum_{k=1}^{M} x_k\right)/(M+1)$, and $j_c = \left(j_b + \sum_{k=1}^{M} y_k\right)/(M+1)$. 4. Players locating in the camera moving direction are more important than players locating in an opposite direction. The camera generally tracks the soccer ball and focuses on interesting content. Let the soccer ball be the pole in the polar co-ordinate system (PCS). The angle of the kth player in the PCS is denoted by θ_k, and $\theta_k = \pi \times \left[1 - (1/2)\mathrm{sign}(y_k - j_b)(1 + \mathrm{sign}(x_k - i_b))\right] + \mathrm{atan}(y_k - j_b)/(x_k - i_b)$, where $\mathrm{sign}(x) = \{1, \text{if } x > 0; 0, \text{if } x = 0; -1, \text{if } x < 0\}$. Let θ_{C} be camera motion direction in the PCS, and $\theta_{\mathrm{C}} = \pi \times [1 - (1/2)\,\mathrm{sign}(my)(1 + \mathrm{sign}(mx))] + \mathrm{atan}(my/mx)$, where $mx = q_4/s$, $my = -q_3/s$, s is zooming factor, $-q_3$ is panning rate, and q_4 is tilting rate, which are estimated using the method in [6]. After that, clue 4 is modelled by the vector $\mathbf{\Phi} = [\Phi_{ball}, \Phi_1, \ldots, \Phi_k, \ldots, \Phi_M]^{\mathrm{T}}$, where Φ_{ball} is 0, and Φ_k is the included angle between θ_k and θ_{C}, i.e. $\Phi_k = \pi \times [1 + \mathrm{sign}((\theta_k - \theta_{\mathrm{C}}) - \pi)] + \mathrm{sign}(\theta_k - \theta_{\mathrm{C}}) \times \|\theta_k - \theta_{\mathrm{C}}\|$.

4 Visual attention value estimation

Simulating human perception, the fuzzy logic inference system (FIS) [7] is introduced to estimate the visual AV of ball and player based on VAFs. The Mamdani-Assilian (MA) model [6] is employed to build our fuzzy 'if-then' system. First, crisp input matrix $\mathbf{X} = [\mathbf{DB},\ \mathbf{DC},\ \mathbf{\Phi}]$ is 'fuzzified' for each object. Three linguistic sets are designed for each element in $\mathbf{X}$, denoted by $\mathrm{L}(\mathbf{DB}) = \{$small, middle, large$\}$, $\mathrm{L}(\mathbf{DC}) = \{$small, middle, large$\}$ and $\mathrm{L}(\mathbf{\Phi}) = \{$small, large$\}$. Meanwhile, the output from the FIS for each object is also designed as a fuzzy set: $\mathrm{L}(\mathbf{O}) = \{$small, middle, large$\}$, which describe the object ability of attracting attention. The membership functions are triangular functions. As shown in Fig. 1, 18 fuzzy rules are designed according to the experiences of watching soccer videos. Through 'defuzzification', 18 fuzzy results are transformed into one crisp output that indicates the AVs of ball and players [7]. We utilise centre of gravity (COG) 'defuzzification' [7] to calculate visual AVs

$$AV_i = \sum_{j=1}^{S} u_i'(y_j) y_j \Big/ \sum_{j=1}^{S} u_i'(y_j),\ i = 1, 2, \ldots, M+1$$

where AV_i represents the visual attention value of the ith object, and $u_i'(y_j)$ is the membership at value y_j in the output distribution of the i th object.

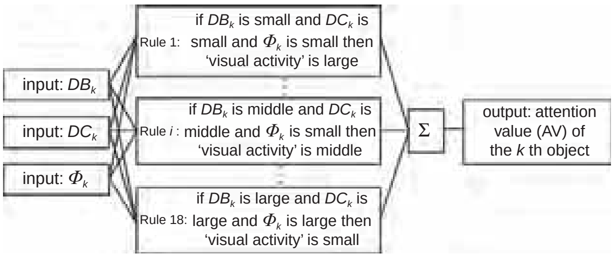
input: DB_k
input: DC_k
input: Φ_k
Rule 1: if DB_k is small and DC_k is small and Φ_k is small then 'visual activity' is large
Rule i : if DB_k is middle and DC_k is middle and Φ_k is small then 'visual activity' is middle
Rule 18: if DB_k is large and DC_k is large and Φ_k is large then 'visual activity' is small
Σ
output: attention value (AV) of the k th object

Figure 1 *Structure and 18 inference rules of FIS*

5 ROI determination

To maximise interesting content within the ROI, two criteria are designed to determine the ROI of each frame: 1. objects with larger AV should have higher priority to be included in the ROI; 2. as many objects as possible should be included in the ROI. An example of ROI determination is shown in Fig. 2. The rectangle indicates determined ROI. Players connected to each other are detected as one object and estimated for one AV, such as number 2. Then the trajectory of the ROI's centre is smoothed in the temporal domain by the polynomial curve fitting model (order is 20 in this Letter) which is an efficient algorithm with low computational complexity. Note that VAF calculation and

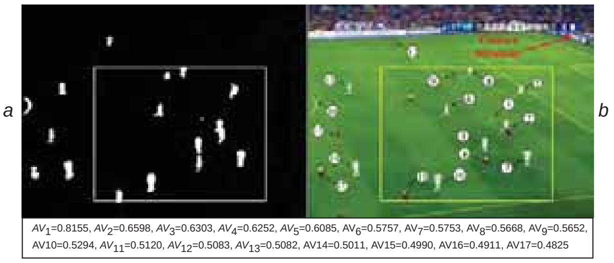
a
b
AV_1=0.8155, AV_2=0.6598, AV_3=0.6303, AV_4=0.6252, AV_5=0.6085, AV_6=0.5757, AV_7=0.5753, AV_8=0.5668, AV_9=0.5652, AV10=0.5294, AV_{11}=0.5120, AV_{12}=0.5083, AV_{13}=0.5082, AV14=0.5011, AV15=0.4990, AV16=0.4911, AV17=0.4825

Figure 2 *Example of ROI determination*

a Binary map with detected ball and players
b Objects with order of AV

© The Institution of Engineering and Technology 2011

Table 1 Evaluation of ROI determination

Video	Numbers of ROI determined long-shot frames	Hit-rate (Seo's [1] method)	Hit-rate (proposed method)
Barcelona vs. Inter1	2256	93.9%	99.5%
Barcelona vs. Inter2	3311	87.7%	96.7%
Inter vs. Roma	3889	90.9%	99.4%
Inter vs. Bayern	2719	83.2%	100%
average	12 175	88.9%	98.9%

'if-then' fuzzy rules (as shown in Fig. 1) secure that the ball is always included in the ROI because of the largest AV of the ball among all objects. It is proved by experiments. After the ROI's temporal smooth, the ROI may not contain the ball in the rare case, as shown in Table 1. The hit-rate of the ball in the ROI is high and satisfactory.

6 Results

To evaluate our proposed system, four 5-minute broadcast soccer video sequences encoded in MPEG-2 with frame size of 720 × 540 (UEFA Champions League Semi-final, 2010; Coppa Italia, 2010; UEFA Champions League Final, 2010; FIFA World Cup Final, 2010) are employed. Both objective and subjective experiments are performed. The adapted video frames, compared with the direct downsampling method and Seo's method [1], are shown in Fig. 3. Here, the targeting frame is 368 × 272. There are 16 random volunteers invited for subjective evaluation. Volunteers are required to answer the following four questions after watching original videos, our retargeted videos, direct down-sampling videos, and videos retargeted by Seo's method. 1. Is our approach better than direct down-sampling? 75% of the answers indicate 'Better', 12.5% of the answers indicate 'No different' and only 12.5% are 'Worse'. 2. Is our approach better than Seo's? 87.5% of the answers are 'Better'; 0% is 'No different'; and 12.5% are 'Worse'. 3. Does the cropping region really represent interesting areas? All volunteers (100%) answer 'Yes'. 4. Are generated retargeting videos reasonable in a temporal consistent way? 87.5% of the answers indicate 'Reasonable' and only 12.5% answers indicate 'Unreasonable'. In objective evaluation, the hit-rate of the ball indicates that the ball is contained in the ROI for most cases (as shown in Table 1).

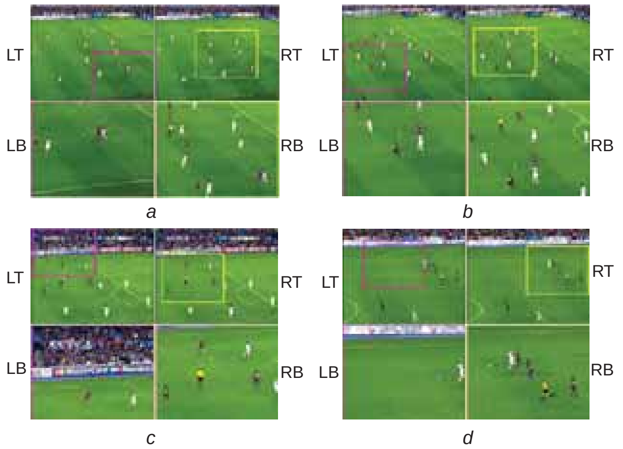

Figure 3 *Four examples (a, b, c, d) of retargeted video frames*

Left top (LT): original frame with ROI window using Seo's method; left bottom (LB): targeted frame using Seo's method; right top (RT): original frame with ROI window using our method; right bottom (RB): targeted frame using our method

7 Conclusion

Experimental results prove that the FIS can well integrate human knowledge to determine the ROI. Content aware ROI determination is proved to be feasible, efficient and satisfactory to users. Visual perception is successfully modelled by VAFs extraction, AV calculation and ROI determination. In future work, other perception clues will be modelled and added into the FIS system. Moreover, this framework will be extended to applications on other video domains, such as broadcast basketball videos, football videos, and tennis videos.

8 Acknowledgment

This work was supported by the National Nature Science Foundation of China (60802073) and UTS Early Career Research Grant.

9 References

[1] SEO K., KO J., AHN I., KIM C.: 'An intelligent display scheme of soccer video on mobile devices', *IEEE Trans. Circuits Syst. Video Technol.*, 2007, **17**, (10), pp. 1395–1401

[2] LU T.R., YUAN Z., HUANG Y., WU D.P., YU H.: 'Video retargeting with nonlinear spatial-temporal saliency fusion'. Int. Conf.

© The Institution of Engineering and Technology 2011

on Image Processing, Hong Kong, September 2010, vol. 1, pp. 1801–1804

[3] CHENG W.H., WANG C.W., WU J.L.: ‘Video adaptation for small display based on content recomposition’, *IEEE Trans. Circuits Syst. Video Technol.*, 2007, **17**, (1), pp. 43–58

[4] GUO Y.W., LIU F., SHI J., ZHOU Z.H., GLEICHER M.: ‘Image retargeting using mesh parametrization’, *IEEE Trans. Multimedia*, 2009, **11**, (5), pp. 856–867

[5] PEI C., GAO L.: ‘A real time ball detection framework for soccer video’. IWSSIP 2009, Challeida, Greece, 2009, pp. 392–395

[6] TAN Y.P., SAUR D.D., KULKARNI S.R., RAMADGE P.J.: ‘Rapid estimation of camera motion from compressed video with application to video annotation’, *IEEE Trans. Circuits Syst. Video Technol.*, 2000, **10**, (1), pp. 133–146

[7] KUNCHEVA L.I.: ‘Fuzzy classifier design’ (Physica-Verlag Heidelberg, New York, 2000)

© The Institution of Engineering and Technology 2011

Interview – Tomas Gonzalez-Sanchez

What is your current role and area of research and what is your background?

I am a PhD student at Rovira i Virgili University (URV) and currently a member of the Intelligent Robotics and Computer Vision Group (IRCV). Since 2005, I have been participating in several research projects where our main goals are focused on improving computer vision and autonomous robot methods.

In 2005 I completed my BS degree in Computer Science Engineering specialising in systems at URV. Then in 2007 I received an MSc in Computer Engineering also from URV. In 2009 I received an Inter-University Master Degree in Artificial Intelligence at URV-UPC-UB and a Graduation Award: 'Best student of the Master Degree in Artificial Intelligence at URV-UPC-UB 2008/2009'. Since then I have been working in robotics and sensor fusion research.

How did you come to work in this area?

In the past I was a member of a couple of RoboCup teams called Spiteam and TeamChaos. Our research was focused in Soccer Standard League with research oriented specifically in the vision field. Within the RoboCup competition, there is a league whose principal aim is to develop robotic technologies for service and assistance, it is called Home League.

In the last few years, many new devices have become available, including Microsoft's *KINECT*, that allow people to interact with computer systems in different ways. So human-robot interaction and co-operation is also currently an interesting area of research where there are a lot of challenges to be addressed in the following years.

What have you achieved in the work reported in your *Electronics Letters* paper?

There have been many attempts over the years to allow humans to communicate with robots in a natural way, and human gestures and signs are the most used methods. Both methods have limitations and must be adapted to be useful for a concrete scenario. Our work was oriented to avoid environmental limitations and to perform robot teleoperation (control from a distance) without any kind of previous technical knowledge about robotics on the part of the user. Our method is aimed to help anybody to control a humanoid robot using natural human gestures.

What applications is this most useful for and what impact do you think it will have?

It will take a long time for the humanoid robots to be used in all kinds of tasks autonomously, given the current limitations of artificial intelligence. On the other hand, teleoperation with many types of robots has been carried out without any problems. So, the possibility to control humanoid robots in a natural way, just with gestures of the body, could allow the use of humanoid robots for a broad number of people without high technical knowledge. Our research can help to make the humanoid robot teleoperation possible sooner.

What have you been doing to develop the system since your Letter?

The Letter was focused on achieving that a robot could reach a set of predefined human gestures and that the algorithm could distinguish a between the gestures. Currently, we are working to deal with dynamic body gestures. The principal challenges are: to detect the human gesture because there are a lot of gestures and to detect when legs and arms are overlapped.

The RGB-depth cameras are usually limited to indoor environments because the camera acquisition sensor has problems with natural light and some kinds of artificial light. So our research is also focused on overcoming that sensor limitation and to determine when the sensor information is wrong because of the environmental light.

What else are you working on at the moment?

I am developing methods to combine RGB-depth acquisition information with other kinds of sensors that provide us with robust information about the environment. This information can be used to improve human–robot interaction and help in labour-intensive human tasks. Sensor fusion is commonly used in the simultaneous location and mapping (SLAM) techniques, where we also have conducted some research. In recent years SLAM using only a simple camera to perform the mapping and navigation, has been an important research field. We are working on performing this kind of SLAM with RGB-depth cameras.

How do you see this field progressing over the next decade? What would you like to see?

Human–robot interaction could become an essential part of life in future. Nowadays, robots have become an active element in our workspace where they are developing tasks in hospitals, museums and many other places. A good example of this future is Robonaut 2 (R2), a NASA robot. R2 is a human-like robot that consists of a head and a torso with two arms and hands, designed to be a permanent resident of the International Space Station. R2 was designed to be a robotic assistant. There is a long way to go before it will be possible to make R2 completely autonomous in its common daily tasks but NASA has made the first steps towards this.

© The Institution of Engineering and Technology 2011

Selected content from *Electronics Letters*

ISSN 2041-5923

Real-time body gesture recognition using depth camera

T. Gonzalez-Sanchez D. Puig

Intelligent Robotics and Computer Vision Group, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans, 26, Tarragona, Spain
E-mail: tomas.gonzalez@urv.cat

Abstract: Human body gesture recognition is a common problem in human–robot interaction. Presented is a novel method for body gesture recognition using a RGB-Depth camera. The proposed technique is able to recognise body gestures, which allow communication between human and robot in order to perform a set of actions by a robot. Also presented is a simple, but fast and effective method for body gesture recognition, which includes illumination invariant skin-colour segmentation.

1 Introduction

Human–robot interaction (HRI) is a multidisciplinary research field, the purpose of which is the study of the interaction between humans and robots. Many HRI interaction approaches have been presented during recent years [1]. There are two main kinds of HRI approaches: (i) those that use electronic devices to communicate with robots, such as touch screens and sensing gloves; (ii) those that use visual information to perform the interaction, such as the methods based on pattern recognition, hand gesture recognition or body gesture recognition.

Walderr *et al.* [2] introduced in 2000 the concept of hand command gesture interface for controlling a wheeled robot being equipped with one manipulator. On the other hand, vision-based techniques can be based on one or a set of cameras in order to obtain maximum data in the minimum elapsed time. Since time-of-flight (TOF) cameras are available, we can combine colour and depth information in real-time to avoid objects located over some specific depth value.

Time-of-flight cameras [3–5] are relatively new acquisition devices that combine a conventional image capture sensor with a new IR-based sensor. The latter is capable of measuring the distances to the objects in the image scene. When both sensors merge the acquired data, the objects in the scene can be characterised by both colour and depth. The information obtained with TOF cameras, also known as RGB-Depth cameras, is comparable to that obtained with stereo-vision cameras but the new sensor performs well in untextured homogeneous scenarios, at a difference of the stereo-vision devices. In that way, we can avoid a training stage dedicated to removing undesired background information. Thus, the method introduced in this Letter for body gesture recognition is fast, effective, and includes illumination invariant skin-colour segmentation.

In previous works, hands tracking algorithms are performed by using HSV colour, RGB colour, or grey-level information. In general, the methods can be classified into not light invariant or light invariant, increasing the computational cost in the segmentation process when the latter techniques are applied.

The use of hidden Markov models (HMMs) to determine human gestures has been common in the literature, such as Park *et al.* [6], with each gesture pose being determined by a five-state graph, or Gaussian hidden Markov models (GHMMs) in Wang *et al.*'s work [7]. Thirteen body gestures were predefined in Park *et al.*'s work. Thus, a gesture recognition method needs, at least, 65 states. Park *et al.*'s work assumed that there is a human static pose that must be reached among two gestures. This static pose helps the gesture recognition algorithm to distinguish two different poses.

© The Institution of Engineering and Technology 2011

2 Proposed algorithm

The proposed human body gesture recognition algorithm consists of three stages: (i) background subtraction, (ii) hands and face detection, and (iii) body gesture recognition.

(i) *Background subtraction:* First, the near and background objects are avoided using a double depth value threshold. Then, the algorithm searches the human body shape using an heuristic function *f(m)* based on the regional area corresponding to each depth level:

$$\text{depth threshold value} = \max\{f(m)\},\quad m = 2.0, 2.1, 2.2, \ldots, 4.0$$

The optimal camera's depth range is from 1 to 4 metres, but the minimum distance needed to acquire an entire body image is 2 metres.

(ii) *Hands and face detection:* Once the body shape is obtained in the previous stage, a fast distance transform method is applied for extracting the body skeleton. The body skeleton extremes are used as location seeds in a region growing algorithm in order to find the regions containing the head and hands. The regions determined in the previous step are compared and selected by a heuristic method to classify regions as: head, hands or unused. Then, a region growing algorithm is applied by using the estimated positions as seeds positions to determine the hands shape accurately. In this step, HSV thresholds for each region obtained by the region growing algorithm are stored and they will be used in the following stages.

In this Letter, we present a new methodology to search those HSV threshold values when the human skeleton is previously determined using a RGB-Depth camera. The entire body detection process is shown in Fig. 1 over one image of the sequence. Once the head (H) and the hands (LH, RH) are detected, the 3D spatial information is extracted from each of them. Then, the human pose is determined by a 9-dimensional vector = {Hx, Hy, Hz, LHx, LHy, LHz, RHx, RHy, RHz}.

(iii) *Body gesture recognition:* The proposed method uses a GHMM in order to recognise human gestures. Similar to [6], the proposed method predefines 13 body gestures. However, differently to Park *et al.*, our method only requires as many states as predefined gestures, since a static pose among two gestures is not necessary. The proposed GHMM-based architecture is shown in Fig. 2 where E_0 is the initial state, and E_i, E_j represent any of the possible gestures.

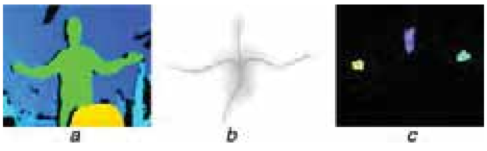

Figure 1 *Gesture recognition process*

a Depth-image
b Skeleton
c Head and hands detection

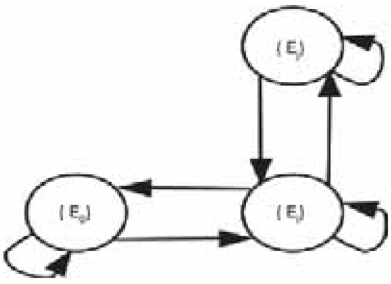

Figure 2 *Architecture of proposed GHMM-based gesture recognition method*

This Letter proposes a new algorithm that allows human gesture recognition without any kind of unnecessary additional gestures. Thus, given a set S = {s_0, s_1, …, s_n} of sequences, the gesture recognition problem can be seen as an evaluation problem, where a gesture G(S) is classified as belonging to a model with a certain probability P(S)

$$G(S) = \arg\max\{P(S)|\lambda_n\},\ n = 0, 1, 2, \ldots, N-1$$

where P(X) is the probability of X, and λ_n is the model of the *n*th gesture. As previously stated, similar to Park *et al.*'s work, the gesture recognition method presented in this Letter considers a set of 13 gestures that have been defined in order to directly compare the results produced by the proposed method with those obtained in the previous related work. Our purpose is also to outperform the classical, stereo-vision algorithms that have the aforementioned well-known problems when the scene is homogeneous in colour and there is not enough texture information to distinguish different objects.

Finally, the BICA architecture introduced by Martin *et al.* [8] is applied to teleoperate the robot by using a server–client model, due to the fact that the depth camera cannot be inserted into the robot architecture.

3 Experimental results

The aforementioned set of 13 gestures was selected from Park *et al.*'s and Wang *et al.*'s sets in order to compare the method presented in this Letter with those two techniques. First, the ability of the method to distinguish between two different human gestures is shown. At this testing phase, a set of 3000 gestures obtained from a sequence of images has been used. Results obtained with the two aforementioned state-of-the-art methods are also included in order to highlight the advantages of the proposed method.

© The Institution of Engineering and Technology 2011

Table 1 Human gesture recognition results using proposed method (tested gestures are same as proposed in [7])

		SR	SB	SL	FB	DB	DL	DB	UR	UL	None
Gesture	3000	315	294	309	298	316	317	319	305	342	185
SR	31.4	307	-	-	-	-	-	-	-	7	-
SB	294	-	294	-	-	-	-	-	-	-	-
SL	308	-	-	304	-	-	-	-	4	-	-
FB	298	-	-	-	298	-	-	-	-	-	-
DB	326	-	-	-	-	309	-	-	-	-	17
DL	314	-	-	-	-	-	312	-	-	-	2
DR	322	-	-	-	-	-	-	318	-	-	4
UR	306	-	-	5	-	-	-	-	301	-	-
UL	343	8	-	-	-	-	-	-	-	335	-
None	175	-	-	-	-	7	5	1	-	-	162

Table 2 Comparison between different gesture recognition methods

	Proposed method	Stereo based [7]	Mono camera [6]
Speed	25 fps	10 fps	30 fps
Accuracy	98%	91.92%	90%
Image	(640 × 480)	(512 × 384)	(640 × 480)
Depth	Yes	Yes	No

Table 1 shows a significant sample of the results produced by the method introduced in this Letter to sort the gestures carried out by a person. The results obtained can be compared with those obtained by previous works, where other acquisition devices were employed (Table 2). To perform an accurate test of the proposed method, four aspects are taken into account in the comparison: algorithm speed, accuracy, image size and depth knowledge. In spite of Wang *et al.*'s work [7] being oriented for gaming applications, it can be directly compared. Wang *et al.* defined a set of nine static body gestures that have the same complexity as Park *et al.*'s gestures [6]. The stereo-based system used in [7] also produces 3D information that implies a reduction of the frame rate to 10 fps. All the methods compared to the technique presented in this Letter give more than 90% of success in the gesture recognition.

The camera used for our experiments reaches a maximum acquisition speed of 25 fps. Then, the images utilised in this comparison were firstly stored, and the algorithm was then applied in order to calculate the time needed to process an image. The processing average per image was 0.032 seconds. Thus, the system proposed can work at 32 fps approximately.

4 Conclusion

Presented is a methodology for human gesture detection with background subtraction using a RGB-Depth camera. Once the body region and the HSV threshold values are found, an unsupervised region-growing-based algorithm is used to locate the shape of the head and the hands. Then, the 3D spatial location of the head and the hands are processed and, finally, human gesture is determined through a clustering algorithm. As a practical application, the recognised body gestures are sent to a Nao humanoid robot that provides a real scenario. In the comparison with previous works, the presented method has the higher recognition ratio situated on 98%. Furthermore, it is shown that the proposed method was the fastest. The background segmentation proposed in this Letter is similar to that proposed in Wang *et al.*'s work, but the depth threshold value is not predefined as in Wang *et al.* In contrast, we define a dynamic threshold value that is calculated using a heuristic method in the set up phase.

5 Acknowledgments

This work was partly supported by the Spanish Government through project Avanza Competitividad I + D + I, TSI-020100-2010-970.

© The Institution of Engineering and Technology 2011

6 References

[1] MURPHY R., NOMURA T., BILLARD A., BURKE J.: 'Human–robot interaction', *IEEE Robot. Autom. Mag.*, 2010, **17**, pp. 85–90

[2] WALDHERR S., THRUN S., ROMERO R.: 'A gesture-based interface for human–robot interaction', *Auton. Robots*, 2000, **9**, (2), pp. 151–173

[3] LARSEN R., BARTH E., KOLB A.: 'Special issue on time-of-flight camera based computer vision', *Comput. Vis. Image Underst.*, 2010, **114**, p. 1317

[4] IONESCU D., IONESCU B., ISLAM S., GADEA C., MCQUIGGAN E.: 'Using depth measuring cameras for a new human computer interaction in augmented virtual reality environments'. Proc. IEEE Int. Conf. VECIMS, Taranto, Italy, 2010

[5] TAKAHASHI M., FUJII M., NAEMURA M., SATOH S.: 'Human gesture recognition using 3.5-dimensional trajectory features for hands-free user interface'. ACM Proc. ARTEMIS, 2010

[6] PARK H., KIM E., JANG S., PARK S., PARK M., KIM H.: 'HMM-based gesture recognition for robot control'. Proc. Int. Conf. on Pattern Recognition and Image Analysis, Bath, UK, 2005, pp. 607–614

[7] WANG Y., YU T., SHI L., LI Z.: 'Using human body gestures as inputs for gaming via depth analysis'. Proc. IEEE Int. Conf. Multimedia and Expo, Hannover, Germany, 2008, pp. 993–996

[8] MARTIN F., AGUERO C., PLAZA J.C., PERDICES E.: 'Humanoid soccer player design' in 'Robot Soccer' (In-Tech, Vukovar, Croatia, 2010)

© The Institution of Engineering and Technology 2011