ABSTRACT

Delivery of 3D immersive entertainment to the home remains a highly challenging problem due to the large amount of data involved, and the need to support a wide variety of different displays. Support of such displays may require different numbers of views, delivered over time varying networks. This requires a delivery scheme featuring scalable compression to adapt to varying network conditions, and error resiliency to overcome disturbing losses in the 3D perception. Audio and video attention models can be used in designing an optimal content-aware compression and transmission scheme, by prioritizing the most visually important areas of the video. This paper gives an overview of a content-aware, scalable multi-view audiovisual entertainment delivery framework. Results are shown to evaluate the kinds of error robustness improvements that could be seen using such a system.

Index Terms—Multi-view video, depth map, multi-channel audio, scalable multi-view video coding, attention modelling, error resilient video coding

1. INTRODUCTION

3D multimedia entertainment has gained great attention, following recent advancements in many parts of the 3D multimedia chain, including acquisition of spatial audio and multi-view video, post-production, compression, transmission, rendering and visualisation. So far, a significant amount of work has been conducted to facilitate and improve the compression and transmission aspects of 3D multi-view video [1-3] in the context of 3D-TV [4] and free-viewpoint video [5]. The multi-view video plus depth coding framework is capable of driving most 3D displays, ranging from stereoscopic to multi-view auto-stereoscopic. It is important to generate the per-view depth information to a sufficient accuracy, sometimes using suitable post-processing techniques, to guarantee perceptually acceptable 3D visual quality.

Compared to 2D simulcast video coding of multi-view video data, applying multi-view coding principles provides up to 30% savings in terms of total bandwidth [2]. Nevertheless, the compression performance is not a reliable indicator of applicability, unless there are a limited number of viewpoints. As the total number of views increases, the increase in the necessary bandwidth increases linearly. Hence, the 3D-TV system framework proposed here considers a flexible scheme, where the broadcasted multi-view material can adapt to changing conditions on the transmission links. The proposed transmission system uses a Peer to Peer (P2P) distribution scheme over IP. This is due to its cost effectiveness and flexibility in scaling its bandwidth under changing constraints. Furthermore, reliable distribution is possible by utilising multi-description coding approaches over the multi-paths. Various works related to P2P video streaming are available in the literature [6, 7].

Scalability of the transmitted content in resolution, quality and region of interest is another key feature of the envisaged 3D multimedia entertainment delivery system. Full scalability of the delivered multi-view video and multi-channel audio content provides easy adaptation to both the network and the application (i.e. 2D/3D display). In the mean time, the exact spatial and temporal alignment of the rendered 3D audio and video scenes comprises another active research field in the proposed delivery framework.

3D video and audio attention models as well as 3D QoE metrics have significant importance in assisting media compression and distribution by introducing content-awareness. Identification of perceptually significant parts of the scene (audio and video objects) makes it possible to utilise prioritisation schemes in scalable compression for optimal bandwidth adaptation. It is therefore inevitable to incorporate the 3D audio and video attention models in the proposed framework.

Section 2 gives an overview of the envisaged scalable and content-aware 3D multimedia entertainment delivery framework. The major functional blocks are explained in detail. Section 3 provides some initial test results with a compression and delivery scheme that is a subset of the proposed framework, to justify some of the concepts. Section 4 gives the concluding remarks.

2. OVERVIEW OF THE PROPOSED 3D MEDIA DELIVERY FRAMEWORK

Figure 1 shows a high level diagram of the proposed scalable and content-aware 3D-multimedia entertainment delivery system. Captured multi-view video and multi-channel audio is post processed for calibration and depth map extraction.
Scalable and multi-description video coding and multi-channel audio coding take place simultaneously and the compressed stream is transmitted over DVB-T and IP.

### 2.1. Depth map extraction and post processing

Extracted depth maps have a direct influence on the perceived visual quality of synthesized views or view pairs. Therefore, they need to be accurate to be rendered as visually pleasant videos, and also easily compressible by block based coders. More accurate depth map extraction algorithms usually comprise iterative computations that result in increased computational demand (e.g. [8]). Furthermore, better precision multi-view depth maps are extracted utilising full scene geometry (all viewpoints) at the expense of increased computational demand [9]. On the other hand, pair-wise stereo matching techniques followed by regularisation steps extract depth maps quicker compared to the other schemes, but up to a certain quality that may not be optimum for compression and view synthesis. Efficient post processing schemes can overcome such difficulties by improving inter-view and intra-view consistencies in depth map sequences [10]. Figure 2 shows a segment from a crudely generated multi-view depth map sequence with several inconsistencies (shown with red circles) and the improvements after processing using an edge, motion and depth range adaptive multi-dimensional median filter. The aim in this framework is to achieve better prediction performance in compression for depth maps and multi-view depth map extraction based on perceptual synthesis quality.

### 2.2. Scalable, Multi-Description Multi-View Coding

To minimise bandwidth usage in a P2P scheme, a codec that exhibits high flexibility in the form of scalable data layers, including temporal and spatial resolution, quality and viewpoint layers, is envisaged for the delivery framework. In the context of 3D multi-view video, some research works have considered utilising scalable layer extraction based on motion/disparity compensated temporal/viewpoint filtering [11], [12]. In this way, the majority of inter-view redundancies are removed in the generated scalable layers. On the other hand, error resiliency is not thoroughly studied. As 3D quality is vulnerable to defects due to the error prone characteristics of IP networks, it is essential to incorporate error resilient features. Multiple-description coding is a suitable tool for P2P multipath, where multiple descriptions are transmitted through different routes between peers. Several multiple description scalable video coding approaches exist in the literature for 2D video [13], [14]. The number of descriptions, rate of each individual description and the level of redundancy in them are controlled to adapt to the requirements of the application and conditions of the delivery network under concern. In [15], a scalable multiple description coding approach is evaluated for stereoscopic 3D video. The authors in [15] have proposed to decompose the scalable colour-plus-depth map stream into even and odd frames to generate multiple descriptions. The envisaged scalable multiple description coding exploits most of the inter-view colour texture correspondences, where the achievable compression efficiency is constrained by the amount of viewpoint scalability. The multiple description streams are selected based on the perceptual importance of the 3D audiovisual content, which necessitates the content analysis prior to compression. Audio and video attention models that define the human reaction to environmental stimuli can be used as a guide in classifying perceptually important features. Hence, the redundancy added to the overall 3D media bit-stream may be optimally distributed over the content.

### 2.3. Audiovisual attention modelling

Audiovisual attention modelling has applications in scalable audio/video coding where the components that are not salient or the user does not give attention to can be coded with a lower resolution or removed completely without decreasing the

---

**Figure 1. The scalable and content-aware multi-view audiovisual entertainment delivery framework**
perceived quality. The human perceptual system considered in this framework consists of the visual and aural systems. The visual system perceives the shape, size, colour, brightness, location and depth of an object, while the aural system perceives the pitch, timbre, loudness and location. There exist various visual attention based video coding schemes that prioritises different video regions and encode accordingly [16], [17]. According to the Weber-Fechner Law, the perceived intensity of the stimulus is related to the logarithm of the physical magnitude of the stimulus. Examples to this law are the brightness on the visual system and the sound intensity on the aural system. Gestalt Laws of proximity, similarity, closure, symmetry, common fate and continuity apply to the perceptual organisation of stimuli [18]. Gestalt theory provides effective criteria in clustering audiovisual objects.

Attention attraction level of each audiovisual cluster can then be found as a function of time and space by fusing the perceived intensities considering also the effect of echoic and iconic memories. The fusion model applies different weights for each stimulus type, which can be determined by parametric modelling techniques on subject responses, gathered using an eye tracker similar to the technique described in [19].

3. INITIAL TEST RESULTS

Initial experiments are conducted to illustrate the potential impact of the concepts explained. Tests are conducted with multi-view plus depth map sequences. Two spatially distant viewpoints of the multi-view test video sequence, called “Akko&Kayo” (VGA resolution), are coded and transmitted for virtual viewpoint synthesis. Two compression schemes are tested using SVC. In the first one, two viewpoints are each encoded with two spatial and two temporal scalable layers and in the second one, in addition to the scalable layers, a redundant bit-stream for the base layer is generated. The redundant bit-stream is generated such that only foreground objects that are visually more significant are redundantly encoded and background objects are encoded once.

The visual attention model is modelled with the relative positions of the objects (depth) in the scene. Several error loss patterns (5%, 10% and 20% packet loss rate) over IP are simulated. The generated redundant bit-stream size is kept within 50% of the main bit-stream. For fair comparison, another redundant bit-stream scheme that encodes the centre region of the viewpoint twice is also utilised. The amount of redundancy in both schemes is the same. The reconstruction quality of the individual viewpoints as well as the synthesised viewpoint that uses the two transmitted viewpoints is evaluated. Figure 3 shows the rate-distortion performance graphs for individual viewpoint reconstruction quality (real camera viewpoint). Figure 4 provides the view synthesis performance with three scalable coding schemes at 10% packet loss rate over IP. The synthesised viewpoint is selected as the midpoint between the two transmitted viewpoints.
It is seen from the graphs that, especially at high packet loss rates, the performance of the scheme that utilises the redundant bit-stream is significantly better than the scheme that doesn’t utilise any kind of redundancy. Similarly, the depth adaptation as a result of visual attention consideration improves the performance further by a margin of 2 dB maximum. Figure 5 shows frame sections from two synthesised videos, using two different redundancy schemes at equal bit-rate and under the same transmission conditions. It can be seen from the figure (especially from the regions with red boxes) that the proposed redundancy scheme based on depth adaptation overcomes many rendering related artefacts in the vicinity of object transitions.

4. CONCLUSION
This paper has given a brief overview of the envisaged multi-view audiovisual entertainment delivery framework. It is crucial to adapt the audiovisual stream according to the network conditions and end-user display capabilities. A fully scalable multi-view video plus depth map stream accompanied with multi-channel audio to be transmitted over a combination of DVB-T and P2P on IP is parsed into multiple descriptions to cope with losses especially on IP transmission links. Perceptually optimised streaming of excessively large audiovisual content over bandwidth limited channels becomes possible for driving next generation 3D media home entertainment systems. Initial test results have shown that prioritizing visually important areas is beneficial for scalable multi-view video transmission.

5. ACKNOWLEDGEMENT
This work has been supported by the DIOMEDES project, funded under the European Commission ICT 7th Framework Programme.

6. REFERENCES