An Innovative No-Reference Metric for Real-time 3D Stereoscopic Video Quality Assessment

Yi Han, *Student Member, IEEE,* Zhenhui Yuan*, *Member, IEEE,* and Gabriel-Miro Muntean, *Member, IEEE*

***Abstract*—Three-dimensional (3D) video technologies have been widely adopted by video service providers and consumer electronics stakeholders due to their potential of offering an immersive user experience. In case of 3D video streaming, the dynamic network conditions are the bottleneck that limits the content delivery at good perceived quality levels and an effective solution is to employ advanced 3D video adaptation schemes. Accurate real-time objective 3D video quality assessment is a critical factor in adaptive decision making. State-of-the-art objective 3D video quality assessment methods are in general reference-based and require the availability of the original 3D video sequence, which makes them not suitable for real-time applications. This paper proposes the extended No reference objective Video Quality Metric (eNVQM), an innovative metric for real-time 3D video quality assessment. eNVQM estimates the 3D video quality by taking as the input parameters network packet loss, video transmission bitrate and frame rate. Based on extensive subjective tests, eNVQM models the impact of network packet loss on 3D video at different bitrates and frame rates on the perceived stereoscopic 3D video quality. The performance of eNVQM is investigated by comparing its results with two state-of-the-art objective video quality metrics: structural similarity index (SSIM) and video quality metric (VQM). Results show that eNVQM maintains similar accuracy level in estimating 3D video quality with the alternative reference-based metrics.**

***Index Terms*—3D video; objective quality assessment; non-reference; stereoscopic**

I. INTRODUCTION

THERE dimensional (3D) video technologies have attracted increasing attention from both industrial content service providers and electronic consumers. The support of the sense of depth significantly enhances the user viewing experience, which is no longer abstract, two dimensional (2D) only, but closer to how reality looks like. The advanced development of image processing, display technologies and video coding (e.g. H.264/AVC, H.264/SVC and Multiview Video Coding (MVC)) has enabled wide deployment of 3D video techniques in various application areas. Recently, the newest version of the HEVC/H.265 standard [1] has added support for 3D video coding, allowing for 3D video encoding with substantially improved video quality at the same bitrate as when using H.264. Additionally, the rapid increasing capacity and speed of networks makes possible the delivery of high definition 3D video to a large user base such as mobile, tablet and wearable devices users. Also, these developments open new revolutionary opportunities for diverse applications beyond the traditional theatre-based 3D movies, such as mobile 3D video streaming, 3D video live chat, 3D conferencing, remote 3D presentation, immersive 3D video gaming, etc. Global organizations have been setup to enhance the academic communication and standardization. For example, 3D@Home [2] focuses on the physiological effects of 3D entertainment, leveraging connections with many nation-wide organizations including China 3D Industry Association [3] and 3DConsortium of Japan [4].

3D video enhances the viewing experience by introducing to users the sense of depth. However, in order to provide users good 3D video quality, there are challenges specific to 3D video, in addition to those that already exist in relation to 2D video. Typical 3D video content consists of views for left and right eyes separately that can be stored in various formats. This 3D video content can be stored in a stereoscopic format [5], which stores two views for left and right eyes, colour plus depth format [6], in which the display terminal uses depth information to recover the two or more views, and a multi-view format [7], which can create multiple views to be viewed from different viewing points [8]. 3D video often has redundant information that can be reduced by various algorithms during the compression process. The sense of depth in 3D video is created by the difference between the views, which may enhance or degrade the overall 3D viewing experience depending on the effect of the image compression/ decompression and delivery.

Fig. 1 illustrates the delivery process of 3D video content which involves capture, transmission and display. In adaptive approaches, there is also a fourth phase which sends feedback from the (dis)play to the capture and/or transmission stages. In these stages diverse devices, equipment and approaches are employed having different requirements in terms of video quality, delivery performance, cost, etc. The capturing device sets the original quality of the video, and its encoding format

This research was supported in part by the Telecommunications Graduate Initiative (TGI), funded by the Irish Higher Education Authority under the Programme for Research in Third-Level Institutions (PRTLI) Cycle 5 and co-funded under the European Regional Development Fund (ERDF).

Y. Han is with the Performance Engineering Laboratory, University College Dublin, Belfield, Dublin 4, and with Performance Engineering Laboratory, School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland (e-mail: yi.han@ucdconnect.ie).

*Z. Yuan is the corresponding author and is with the Ministry of Education Key Frequency and Circuits Lab, School of Electronic and Information, Hangzhou Dianzi University, China (email: yuanzhenhui@hdu.dedu.cn).

G.-M. Muntean is with the Performance Engineering Laboratory, School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland (e-mail: gabriel.muntean@dcu.ie).

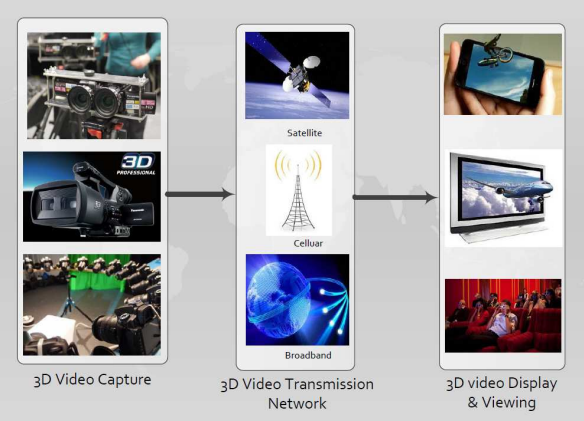


Fig. 1. 3D Video Delivery Process

and settings influence the efficiency of both the storage and transmission processes. The major approaches in the capture process are stereoscopic, colour-plus-depth, and multiview 3D video-based, capturing the 3D scene from one point of view or from different perspectives, in the latter case [9]. The same approaches must be employed to reproduce the 3D scene in the display process at the viewer. Although very interesting, analysing the 3D content capture and display processes or considering the multiview 3D approach are not in the scope of this paper. Particularly, this paper focuses on studying the manner in which the 3D video quality is affected in the transmission process of stereoscopic 3D video. In the 3D video delivery, network impairments that affect either view (left or right) may result in different level of degradation of the overall 3D video quality. Additionally, encoding at different bitrate and frame rates may have different impact on the 3D video than that on the 2D video.

Network delivery of 3D video content at good quality levels is challenging mostly due to highly dynamic network conditions. The delivery performance is affected by network-induced impairments, especially for mobile and real-time interactive applications. Adaptive delivery schemes [10]-[13] in 2D video have been proposed by various researchers to monitor network environmental changes and adjust dynamically the video delivery settings (e.g encoding parameters, buffer size, etc.) These adaptive solutions require knowledge of current 2D video Quality of Experience (QoE) estimates which are obtained from using objective 2D video quality metrics. ITU-T G.1070 [14] defines standardized objective 2D video quality metric for estimating 2D video quality.

However there is a lack of such accurate metrics for estimating 3D video quality which can be used in adaptive 3D video transmissions. Several researchers have focused on assessing the QoE of 3D video and they have used both subjective and objective quality assessment methods. Subjective methods (i.e. involving people evaluating the video quality) provide highly accurate results in terms of video quality that directly reflect human perception of the quality levels. However, these methods require carefully controlled environments with least impact factors such as background noise, light condition, room size, equipment, etc. Furthermore they are time consuming and human resource intensive and thus are not suitable for real-time assessment during transmission. Objective methods have lower accuracy but they are preferred as they can be conducted during the transmission. Several objective 3D video quality metrics have been proposed recently [15]-[20]. However the lack of accuracy in these metrics is mainly due to the fact that the human visual system (HVS) is difficult to model in 3D by analysing pixels and depth as they are reference/content based. Other factors that affect HVS include eye comfort level, viewing distance, luminance, etc. The widely used 3D video quality methods employ 2D video quality metrics, including PSNR [21], SSIM [22], and VQM [23]. The quality of left and right views of 3D video are evaluated separately and averaged by different weights to an overall 3D video quality [15] [16]. These methods require the original and degraded video sequences in order to analyse the blockiness, blurring effect, and depth information of the videos by modelling HVS. Authors of [24] proposed joint bit allocation and rate control for coding multi-view 3D video, based intrusive methods to calculate the view synthesis distortion from original and generated view in colour-plus-depth 3D video. Furthermore, the existing 3D video quality assessment methods are highly dependent on the video content and do not consider the effect of network delivery-induced impairments. These quality metrics can only be used when both the original and received video sequences are available, after the transmission and therefore they are not suitable for real-time adaptive transmissions. The no reference PSNR [25] for 2D video can be used in real-time, but the additional depth sense cannot be reflected by simply averaging the quality of the left and right views.

This paper investigates the effect of network delivery condition variations on the 3D video quality by considering diverse content with different video bitrates and frame rates. The **extended No reference 3D Video Quality Metric (eNVQM)** for stereoscopic 3D video quality assessment is then proposed. Employing the philosophy behind the ITU-T G.1070 model for 2D video quality assessment [14], eNVQM proposes a new model for 3D video quality based on the results of subjective tests which assess 3D video user perceived video quality including eye comfort level, enjoyment, and quality of experience enhancement. eNVQM is derived from the correlation between network packet loss rates and 3D video quality for different combinations of bitrates and frame rates.. eNVQM estimates the 3D video quality in real-time during the transmission and can be used for proactive adaptation in 3D video streaming. eNVQM extends a previously proposed metric [26][27] which considered a single fixed bitrate and frame rate only.

The rest of the paper is organized as follows. Section II presents the state-of-the-art 3D video quality assessment methods. Section III describes the mathematical model of eNVQM in details. Section IV explains the derivation of the eNVQM model through experiments. Section V analyses the experimental results and performs comparison against other 3D video quality metrics. At the end, conclusions are drawn and future work directions are indicated in Section VI.

II. TECHNICAL BACKGROUND AND RELATED WORKS

Stereoscopic 3D video creates or enhances the illusion of depth in an image by presenting two offset 2D images to the left and right eye of the viewer, respectively. These two images, representing two perspectives of the same object or scene (also called *views*, e.g. left/right view), are then combined in the human brain to provide the perception of 3D depth. The 3D depth sense is produced as a result of a minor deviation of the two views similar to the perspectives that both eyes perceive in natural binocular vision. This is considered as the easiest way to enhance depth perception in the brain in comparison with other methods [28]. The stereoscopic 3D for broadcasting has been discussed in Rec. ITU-R BT.1198 [29] as one of the earliest recommendations for this format of 3D video. The document has proposed that a stereoscopic broadcasting system based on right and left eyes should not cause significant problems (such as eye-fatigue, "puppet theatre" effect, etc.) and should not provide lower quality than traditional SDTV systems. It also recommends that the stereoscopic system should be maximally compatible with monoscopic TV broadcasting systems.

Subjective assessment methods for video have been proposed by ITU in ITU-R BT.500 [30] for television pictures, ITU-T P.913 [31] for Internet video as well as distribution quality television in any environment, ITU-T P.3D-sam [32] specifically for 3D video quality, ITU-T J.3D-fatigue [33] for 3D video visual fatigue and safety guideline assessment. Additionally the 3D display requirements when conducting subjective quality assessment tests have been specified in ITU-T J.3D-disp-req [34] in details. If autostereoscopic displays (ASD) are used, the ISO/TR 924-331 standard, which establishes ergonomic optical requirements aiming of reducing visual fatigue caused by stereoscopic images on ASDs, is highly relevant. This standard also proposed performance characteristics to evaluate various aspects of 3D viewing experiences, such as 3D crosstalk, interocular luminance difference, interocular chromaticity difference, etc.

The principle of such 3D video format allows for simple creation of 3D content and no or little additional image processing is required. The stereoscopic 3D format can be side-by-side (SBS) or top-bottom, representing the layout of the two views in the 3D content. While transmitting over the network, the two views are combined into a 3D stream, in which the left and right view frames are following each other in sequential manner. For example, in the case of Blu-ray content, the video is encoded and sent out from the sender at 720p at 24 frames per second, per view, or 48 frames per second. It is stored frame by frame interleaving left and right views in a sequential manner. Thus when a packet is lost, either the relevant left or right view is affected. The technique is described in details in [35].

Various methods have been investigated to assess the 3D video quality. Authors of [15] studied the performance of assessing stereoscopic 3D video quality under various packet loss scenarios using 2D objective quality metrics, including PSNR, SSIM and VQM. They averaged the results for the left and right views of the 3D video, and showed that when using PSRN and SSIM better correlated results with the 3D video depth perception are obtained than when VQM is employed. Another study [36] using a similar method showed that the colour component is dominant in the overall 3D video quality perception, while depth has less impact. The quality assessment of colour plus depth based 3D video using these 2D video quality metrics is described in [16], in which the left and right views are rendered using Depth-Image-Based Rendering (DIBR) technique. Another method considers 1/3 and 2/3 weights for left and right views respectively when using PSNR to evaluate the two views [18], but this split seems arbitrary.

Apart from using 2D video quality metrics, new metrics are also proposed for 3D video quality assessment. A crosstalk perception assessment method for stereoscopic 3D video is described in [37]. The crosstalk perception is considered as a result of shadow degree, separation distance, and spatial position, which happen in the visualization stage of stereoscopic imaging. However the overall 3D video quality perception is also affected by various factors. Perceptual Quality Metric (PQM) [20] is more sensitive to pixel level image degradation and error quantification than when these happen at sequence level. Authors of [38] propose an objective model that predicts the quality of lost frames in 3D video streams based on the estimated lost frame size only. A solution which modelled the impact of eye dominance on the perceived 3D video quality by chopping the images into small 4 x 4 blocks based on spatial frequency was presented in [17].

A lightweight no-reference method to estimate the colour plus depth 3D video quality from depth streams using different set of packet layer parameters that are abstracted from packet headers was proposed in [39]. The results presented showed high correlation to SSIM results, but no comparison with subjective test results was given. Also all the video clips used had frame rates of 25 or 30 and a very limited bitrate range only. The exact model parameters were not provided, so no independent validation of the results published can be done.

The 3D video quality assessment methods used in practice have different accuracy levels, as well as diverse advantages and limitations. More importantly, most of them require referencing to the original video source, unlike our proposed eNVQM, which does not require the presence of the original 3D video sequences, enabling it to be applicable to a much larger range of usage scenarios.

ITU-T G.1070 provides a good methodology for mapping bitrate, frame rate and packet loss to the 2D video quality expressed in Mean Opinion Score (MOS). MOS assesses the media quality by its absolute value using absolute category rating (ACR) and evaluates the quality perceived by the viewer, with no reference.

In this paper MOS is employed as it uses absolute rating which closer to the situation in which viewers consume regularly video content in their daily life, having no video reference to compare against when they perform their quality assessment. Additionally, in our methodology the subjective test results have been compared against objective test results, which use SSIM, VQM and ITU-T G.1070 2D video quality metrics, and mapping between SSIM and VQM results to MOS is done easier using existing mapping solutions [41] than if CMOS was adopted.

III. PROPOSED 3D VIDEO QUALITY MODEL

eNVQM models the relationship between network packet loss, 3D video bitrate and frame rate and the 3D video quality. The model takes the above three variables as input and calculates the estimated 3D video quality as output. eNVQM builds on the idea introduced by ITU-T G.1070 [10], which has defined a model for 2D video quality estimation, and introduces depth perceptual quality in modelling stereoscopic 3D video quality. The relationship between colour and depth and the video perceptual quality is calibrated by three factors including eye comfort level, degree of enjoyment, and enhancement of user quality of experience level.

In ITU.T G.1070, the end-to-end delay is considered in the audio quality metric, but not in the video quality metric. This is as the delay has a more important effect on remote audio delivery than on video, which tolerates better larger delay and delay variations. For similar reasons, delay is not taken directly into account in our proposed 3D video quality metric eNVQM, either.

A. ITU-T G.1070 2D Video Quality Metric

The ITU-T has standardized a user opinion model for 2D video-telephony applications in G.1070. It estimates the 2D video quality in telephony applications by considering the network impairment parameters (i.e. packet loss in video) and encoding parameters, including codec type, video format, key frame interval, and video display size.

The ITU-T 2D video quality is modeled by equation (1):

$$V_q = 1 + I_{coding} \, e^{-\frac{Ppl_V}{D_{PplV}}} \quad (1)$$

where Ppl_V represents packet loss rate, D_{PplV} expresses the degree of video quality robustness due to packet loss, and I_{coding} calculates the basic video quality affected the coding impairment that is introduced by video bitrate (Br_V is expressed in kbps) and video frame rate (Fr_V is measured in fps). Note (1+ I_{coding}) represents the video quality when the packet loss is 0%. I_{coding} is calculated as in equation (2):

$$I_{coding} = I_{Ofr} * e^{-\frac{(\ln(Fr_V) - \ln(O_{fr}))^2}{2 * D_{FrV}^2}} \quad (2)$$

Parameter O_{fr} represents the optimal video frame rate corresponding to the video bitrate (Br_V) for the best video quality. It is expressed in equation (3):

$$O_{fr} = v_1 + v_2 * Br_v, \; 1 \le O_{fr} \le 30 \quad (3)$$

If $Fr_V = O_{fr}$, then $I_{coding} = I_{Ofr}$. I_{Ofr} is the maximum video quality of the video at bitrate Br_V and is calculated as in equation (4):

$$I_{Ofr} = v_3 - \frac{v_3}{1 + (\frac{Br_V}{v_4})^{v_5}}, \; 0 \le I_{Ofr} \le 4 \quad (4)$$

In equation (2), D_{FrV} represents the degree of video quality robustness introduced by frame rate (Fr_V) and is calculated using equation (5):

$$D_{FrV} = v_6 + v_7 * Br_V, 0 < D_{FrV} \quad (5)$$

At last in equation (1), D_{PplV} represents the degree of video quality robustness due to packet loss rate and is calculated according to equation (6):

$$D_{PplV} = v_{10} + v_{11} * e^{-\frac{Fr_V}{v_8}} + v_{12} * e^{-\frac{Br_V}{v_9}}, 0 < D_{PplV} \quad (6)$$

In the above equations, v_1, v_2,…, v_{12} are derived from subjective 2D video tests and are dependent on the video coding, and display size. The recommendation gives five sets of coefficients for different display sizes for MPEG-4 and ITU-T H.264, respectively. The methodology for deriving the coefficients in the model is given in [10]. In the standard, the related accuracy of the predicted video quality was evaluated by the Pearson product-moment correlation [42].

The derivation of the proposed eNVQM for 3D video quality assessment is shown in the next subsection.

B. Extended No Reference 3D Video Quality Metric (eNVQM)

The stereoscopic 3D video consists of two views that can be in either left/right or top/bottom format and directed to viewer's left and right eyes respectively, by making use of various display technologies. As there is no difference in terms of viewing experience with the two formats in stereoscopic 3D video, for simplicity of explanation, this paper refers to the left/right format for stereoscopic 3D videos only. The two views are slightly different from each other as they are shot from two close, but different points of view. The two views are then synchronized, displayed simultaneously and the human brain creates a 3D illusion effect from the disparity of the two views, providing the human observer with the sense of depth in the 3D scene. When considering the transmission of such 3D content, the information lost in one view may result in an impaired overall 3D displayed frame and thus in decreased 3D video quality, despite the potentially excellent reception of the other view. For this reason, we believe that network impairment has different impact on 3D video than on 2D video.

Following the same methodology employed in ITU-T G.1070 for mapping bitrate, frame rate and packet loss to the video quality, we propose for eNVQM the formulas presented in equations (7)-(10), where the 3D video quality is expressed by $V^{3D}{}_q$ in terms of MOS.

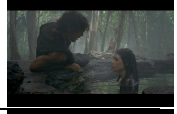
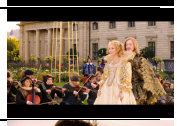
In eNVQM, $I^{3D}{}_{coding}$ is composed of two additive natural logarithm components for both frame rate and bitrate, respectively, reflecting their effect on the video quality when packet loss ($PplV$) is 0%. The exponential component of the eNVQM formula describes the effect of packet loss on the video quality when considering 3D video frame rate and bitrate.

$$V^{3D}_{colour/depth} = 1 + I^{3D}{}_{coding} \, e^{-\frac{Ppl_V}{D^{3D}{}_{PplV}}} \quad (7)$$

$$I^{3D}{}_{coding} = a_1 \ln(Fr_V) + a_2 \ln(a_3 + a_4 Br_V) \quad (8)$$

$$D^{3D}{}_{PplV} = a_5 + a_6 * e^{-\frac{Fr_V}{a_7}} + a_8 * e^{-\frac{Br_V}{a_9}} \quad (9)$$

TABLE I
VIDEO SAMPLES

Clip no.	Motion complexity level	Content scenario	Duration (seconds)	Sample frame
1	High	Running	9	
2	High	Driving	14	
3	Medium	Swimming	13	
4	Medium	Dancing	6	
5	Low	Kissing	8	

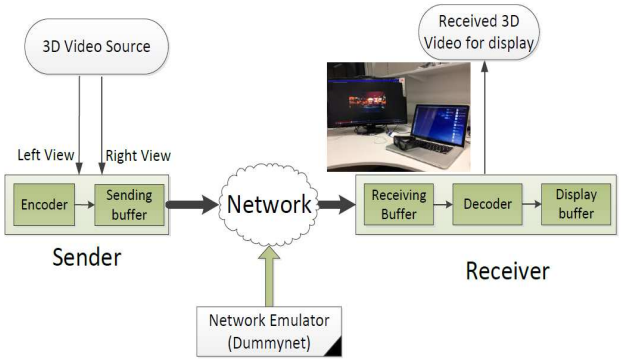


Fig. 2. Experiment Framework

TABLE II
COEFFICIENTS COMPUTED FOR ENVQM

	colour	*depth*
a_1	0.09136	0.08751
a_2	1.11132	1.05853
a_3	0.93128	0.93067
a_4	1.79391	1.7921
a_5	-1.24607	-0.46754
a_6	0.01436	1.67570
a_7	-33.775	-33.03
a_8	2.17023	0.39725
a_9	-5.37876	-4.45855

Equations (7)-(9) are used for quality computation of both colour and depth components of the 3D video: $V^{3D}{}_{color}$ and $V^{3D}{}_{depth}$. Two sets of coefficients A = {a_1, a_2,…, a_9} are derived from subjective 3D video tests involving colour (A_{colour}) and depth (A_{depth}) perception, respectively. a_1 and a_2 reflect the effect of frame rate and bitrate, respectively when there is no packet loss. a_3 and a_4 quantify the contribution of bitrate so that both frame rate and bitrate can be represented in balanced manner in the overall formula. There is no need for frame rate to have similar coefficients to bitrate because of the scale difference between the frame rate range (10 ~ 60 fps) and that of bitrate (1~10 Mbps). The coefficients a_5 to a_9 are used to map different scales of frame rate and bitrate on the scale of packet loss rate, respectively. Coefficients a_1 through a_9 are dependent on the codec type, video format, and display size.

Furthermore, unlike the case of 2D video quality, the overall 3D video quality modelling considers colour and depth perceptual quality, expressed in equation (10):

$$V^{3D}{}_{q} = xV^{3D}{}_{colour} + yV^{3D}{}_{depth}, \qquad x + y = 1 \qquad (10)$$

where x and y are different weights for colour and depth perceptual quality, respectively. It is assumed that there is an additive effect of depth perception on the colour perception in terms of the 3D video quality based on the findings that viewing video content in 3D increases the perceived image quality [15] and depth has a positive effect on visual experience from an enhanced sense of presence [49]. As the sum of x and y is always 1, $V^{3D}{}_{q}$ has the same range as $V^{3D}{}_{colour}$ and $V^{3D}{}_{depth}$. The values for x and y are determined by considering the correlation with three other perceptual factors collected in the subjective tests, reflecting eye comfort level, degree of enjoyment, and level of user quality of experience enhancement.

$V^{3D}{}_{color}$ and $V^{3D}{}_{depth}$, representing the 3D image quality and depth perceptual quality, can be used individually as two quality indicators using equation (7). The overall 3D video quality is estimated by $V^{3D}{}_{q}$ using equation (10), as a joint result of colour and depth perceptual qualities, as well as considering the other three aspects of user perceptual experience.

IV. EXPERIMENTAL SETUP

An extensive set of experiments are conducted to study the relationship between the perceived 3D video quality, network characteristics (i.e. packet loss), 3D video encoding settings (i.e. frame rate, bitrate), for diverse video content. Different network delivery scenarios are considered with a range of network packet loss rates. In order to reduce the dependence on video content, a wide range of 3D video samples with different content types is selected. These videos are then encoded with diverse settings.

Table I shows the five selected video clips, each with content belonging to a different scenario with diverse motion complexity levels. The durations of the selected video clips vary between 6 to 14 seconds, in the range recommended by ITU-T R. P.913 [43] and ITU-T P.3D-sam [32]. These video clips are H.264 /MPEG-4 AVC encoded at high (4 Mbps), medium (3 Mbps), and low (2 Mbps) average bitrates, follow the IPPP MPEG sequence format and have frame rates of 11 fps

and 18 fps, targeting mobile applications. The resolutions of all video clips are 1280 x 720 pixels.

Standard H.264 encoding was used and frame-copy was adopted for error concealment. The use of multiple video sources with different degrees of motion contributes to metric's validation independent from content. Details of the display equipment and 3D display technology were given for the purpose of reference.

The test topology of the experiments is shown in Fig. 2. Two VLC player instances running on two machines are used for sending and receiving 3D stream, respectively. At one end, the built-in x264 library of the VLC tool is used for encoding video streams into the H.264/MPEG-4 AVC format for stereoscopic 3D videos. At the other end, the 3D video stream is captured and decoded into sequence pairs of left and right views in the 4:2:0 YUV format, which is the same as in the original video. During the transmission over the network, *Dummynet* [44] is used to control the desired packet loss rate in the network. The simulated packet loss follows a uniform distribution. *Wireshark* is used at the receiver side to monitor the stream and calculate the packet loss rate. 11 network loss scenarios are created: 0%, 0.1%, 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 8% and 10%. More scenarios were studied in the lower packet loss range (less than 5%) to allow for better study accuracy. Overall there are 11 packet loss scenarios, 3 encoding bitrates, 2 frame rates and 5 different video content types, resulting in 330 video clip left-right pairs transmitted during the experiment. These video pairs are used firstly in subjective tests and then in comparison-based verification when using other objective quality models, as described in details in the next section.

Subjective tests are conducted with 50 volunteers with diverse ages, genders and backgrounds. The 330 video pairs are divided into 10 groups, each containing 33 videos randomly selected with different video content, packet loss, bitrate and framerate. In order to avoid boredom, a time limit of maximum 30 minutes was imposed for each participant. Each group is shown to 4 participants and in this way, each individual 3D video pair of views has at least 4 results from 4 different observers. Considering the five different video content types, each combination of packet loss, encoding bitrate and frame rate is tested 4*5=20 times. In this way, a good balance between the number of subjects testing any individual sample and the total number of tests is maintained. The clips are displayed on a machine with a 27 inches *3D Asus VG278* monitor with resolution 1920 x 1080 pixels, and the *3D vision*^{2} support enabled from *Nvidia.* The 3D player synchronizes and displays the pair of left and right view clips simultaneously. The participants are required to wear a pair of *3D vision*^{2} wireless active shutter glasses in order to watch the 3D effect of the video.

As suggested by the monitor manufacturer, the viewing distance is set to 1 m. All other test setup details follow the recommendations of ITU-T R. P.913 [43]. The tests are conducted in a 5m x 5m quiet room, having the monitor away from windows to avoid additional unnecessary influence of light. Each participant is asked to assess their colour image 3D experience, 3D depth experience, eye comfort level, 3D level of enjoyment, and state how the 3D effect enhances their overall viewing experience. The grading uses the 1 (bad) to 5 (excellent) MOS scale. These will be used for deriving the values of eNVQM coefficients.

V. Result Analysis

The subjective test results consist of grading marks for 330 video clip pairs, each having a particular combination of bitrate, frame rate, video content and packet loss rate. During the subjective tests, each participant has graded five different aspects of the 3D video for each of the 33 videos in a group out of the total number of 10 groups.

Next the eNVQM coefficients are derived according to the grades of the five different aspects of 3D video quality assessment mentioned above.

A. eNVQM Metric Derivation

Among the five aspects, colour and depth 3D quality perception are used to derive coefficients $A = \{a_1, a_2, \ldots, a_9\}$ for colour and depth models, respectively. As the tested videos have different combination of bitrate, frame rate and packet loss rate, the aim is to best map eNVQM to each of these estimations of user perceptual 3D video quality. 25% of the total data had been reserved and used for data validation using holdout validation. The initial data from different subjective tests with different content, bitrate, frame rate and packet loss rate was chosen randomly to 'derivation data set' d0 and 'validation data set' d1. Thus the two data sets contain similar percentages of data with different properties.

The subjective results are carefully processed in order to eliminate outliers introduced by observers. When considering the results for each clip, an outlier is identified if it is scored more than 2 grades adrift from the median MOS of the values recorded from all its viewers. However, when considering packet loss scenarios, for each case there are 22 clips (75% of 30 clips) with different content, bitrates and frame rates (with the packet loss rate fixed) and the highest and lowest 10% of them are considered outliers and are removed. The same process is performed for both overall colour and depth perception, respectively.

The coefficients a_1, to a_9 are calculated for A_{color} and A_{depth} following the steps described in ITU-T G.1070. The method involves calculating some of coefficients by having only one of them variable and keeping the other ones fixed. The coefficients are approximated using the Least Square Approximation (LSA) method [45]. The corresponding fitting curves for both colour and depth parameters are shown in Fig. 3 and Fig. 4, respectively. The corresponding coefficients for colour and depth models instantiated from equations (7)-(9) are presented in Table II.

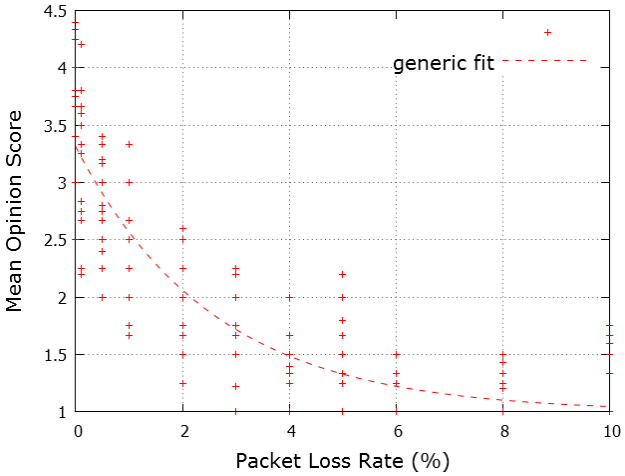


Fig. 3. Mapping colour quality perception vs. network packet loss rate

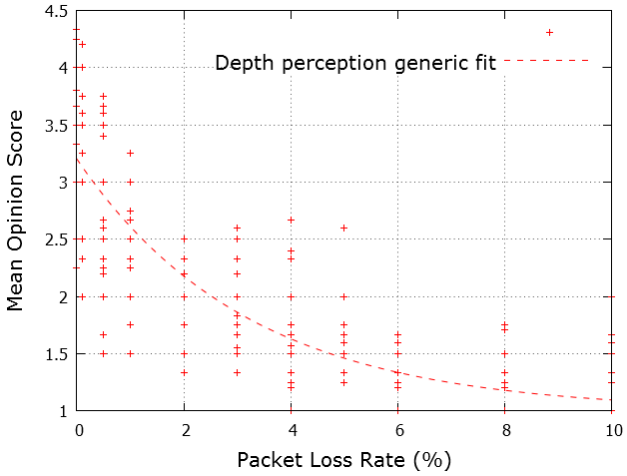


Fig. 4. Mapping depth perception vs. network packet loss rate

In order to verify the correctness of the model, the remaining 25% of the subjective results are used to compute the Pearson correlation with the eNVQM results. The model uses inputs with the same frame rate, bitrate and packet loss rate as in the clips presented to the observers. The correlation results are shown in Table III. Note that high correlation values of 87.3%, 91.6% and 94.2% for colour and 78.5%, 90.3% and 93.5% for depth when 25%, 75% and 100% of the results are considered, respectively, are obtained. Slightly lower correlation values for the 25% of results case are due to the combined effect of the lower number of results considered and their discrete nature (on the 1-5 scale). However the general high level of correlation indicates that our derived model coefficients are valid and reliable.

Next, the weights for colour and depth components are determined by making use of three additional set of results in the 3D video quality assessment regarding eye comfort level, 3D enjoyment level, and 3D effect enhancement level. The same process of removing outliers for each clip was followed, but outliers when considering a particular packet loss rate are retained, as no fitting curve was required to be identified in this step. Giving different weights to colour and depth, the overall

TABLE III
VERIFICATION OF ENVQM COEFFICIENTS - PEARSON CORRELATION OF ENVQM AND SUBJECTIVE RESULTS

	25% subjective results		*75% subjective results*		*100% subjective results*	
	colour	depth	colour	depth	colour	depth
Correlation	0.873	0.785	0.916	0.903	0.942	0.935

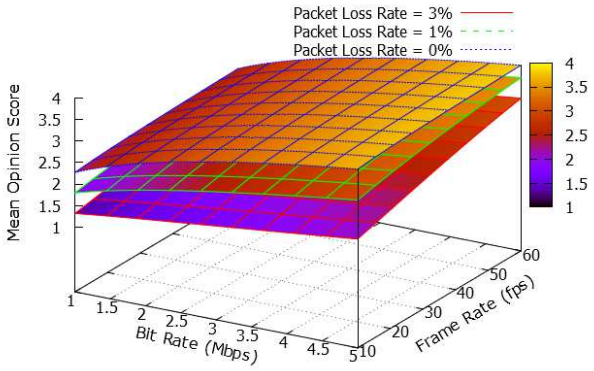


Fig. 5. eNVQM 3D video quality with 0% 1% and 3% packet loss, respectively

scores are compared against the results of the above three factors. Each result set organizes data for a particular packet loss rate in each row with a combination of bitrate and frame values in each column. Correlations are computed for each column pairs containing subjective results and grading marks for the above factors. Finally the average correlations over all packet loss rates are calculated. This is done for each of the three subjective factors considered. The highest average correlation of these factors is considered to determine the weights of x, y for colour and depth perception. The correlation trend follows a 2nd order polynomial function, in which y is replaced by (1-x):

$$Corr = -0.0026x^2 + 0.0046x + 0.8644 \qquad (11)$$

The function of the correlation trend is a parabola of x (since y = (1-x) and its vertex is at x=0.885, giving the highest correlation of 0.866434615. Thus equation (11) can be expressed as:

$$V^{3D}{}_q = 0.885 \times V^{3D}{}_{colour} + 0.115 \times V^{3D}{}_{depth} \qquad (12)$$

where $V^{3D}{}_{\text{colour}}$ and $V^{3D}{}_{depth}$ are calculated using equations (7)-(9) and the coefficients from Table II. The smaller weight derived for depth perceptual quality matches the findings that depth perception plays a less important role in the overall 3D quality than that of color [50].

The eNVQM model takes three input variables: frame rate, bitrate and packet loss rate. The output of eNVQM is expressed in terms of MOS and refers to the human perception of 3D video quality. Fig. 5 illustrates eNVQM variation against bitrate and framerate when the packet loss is 0%, 1% and 3%,

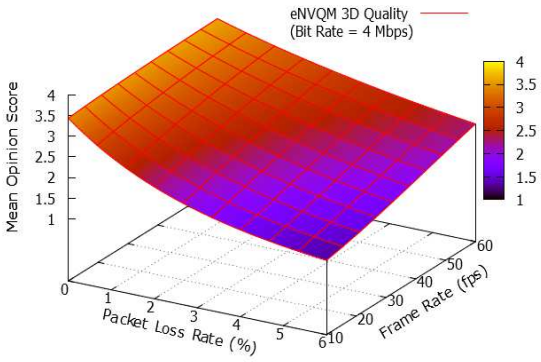


Fig. 6. eNVQM 3D video quality with bitrate 4 Mbps

respectively. It can be noted how MOS increases as bitrate and frame rate become larger and how the effect of bitrate growth is larger in terms of MOS when frame rate increases. For example, at packet loss of 1% and frame rate of 30 fps, MOS is 1.698 for a bitrate of 1 Mbps; MOS increases to 2.186 for a bitrate of 2 Mbps, it reaches 2.60 when the bitrate is 3 Mbps and further becomes 3.26 for a bitrate of 5 Mbps. On the other hand, at the packet loss of 1% and bitrate of 2 Mbps, MOS is 2.12 for a frame rate of 10 fps, 2.16 for 20 fps, 2.18 for 30 fps and only 2.239 for 60 fps. Note the effect of bitrate and frame rate on quality differs for different packet loss rates as shown in different layers illustrated in Fig. 5.

Fig. 6 shows specifically eNVQM variation against loss rate and frame rate at a fixed bitrate of 4 Mbps. It is interesting to see that for lower range frame rates, MOS drops more rapidly relative to packet loss growth, while MOS drops smoothly for higher range frame rates. From eNVQM we can conclude that the encoding bitrate has a higher impact on the overall 3D video perceptual quality than frame rate, and packet loss has a higher impact on the overall 3D video quality when the frame rate is low.

B. Comparison with Other Metrics

SSIM and VQM are the two widely used methods for objective video quality assessment, which were designed to evaluate 2D video quality. They evaluate the 2D video quality by intrusively comparing the original and degraded video samples. Despite our initial reluctance regarding the use of 2D metrics to assess 3D video quality, in order to compare the performance of the proposed eNVQM to other models used in the literature by other researchers, SSIM and VQM were used for 3D video quality estimation. This is as the authors of [15] have shown that SSIM and VQM ratings of average left and right videos can be used as good objective quality models for prediction of 3D perceived quality under packet loss scenarios. MSU VQMT [46] was used as computational tool. Since SSIM and VQM use different scales from MOS, normalization methods described in [47] and [48] were employed, respectively. The original and degraded sample pairs were compared by MSU VQMT for the left and right views, and the average scores of both views converted to MOS scale were compared with the results of eNVQM.

TABLE IV
DIFFERENT METRIC PERFORMANCE COMPARISON

Method	*SSIM*	*VQM*	*eNVQM*
Pearson Correlation	0.911	0.932	0.872
Spearman Rank Correlation	0.851	0.871	0.883
RMSE	1.126	0.329	0.505

Pearson correlation, Spearman Rank correlation and root mean square error (RMSE) were computed comparing the results when using the proposed eNVQM with those when employing existing metrics SSIM and VQM. The correlation testing was performed on the remaining 25% of the subjective testing results, not used in the model building process, ensuring independent model validation. The results are listed in Table IV. These comparative performance results show that by using eNVQM similar accuracy level in predicting the perceived 3D video quality can be obtained with the case when the other two reference-based metrics were employed. For instance when considering Spearman Rank correlation, eNVQM even slightly outperforms both alternative solutions with a result of 0.883 in comparison with 0.871 and 0.851 of VQM and SSIM, respectively.

VI. CONCLUSIONS AND FUTURE WORK

This paper proposes the extended no reference objective video quality metric (eNVQM) for the assessment of stereoscopic 3D video quality during network-based content transmission. Following a methodology similar with that of ITU-T G.1070, eNVQM estimates the 3D image quality and depth perceptual quality using encoding frame rate, bitrate and network packet loss rate and finds an additive effect of the two while also considering three other experience factors. Perceptual tests were performed and their results were employed to both derive parameters for the proposed eNVQM model and test its validity. Statistical results show that eNVQM has similar level of accuracy in terms of human perception of 3D video, in comparison with SSIM and VQM, two commonly used assessment methods. eNVQM can be used for adaptive 3D video transmissions as it can quickly estimate the current video quality so that delivery adjustment actions can be taken at the earliest possible point, increasing user perceived quality levels.

Future work will consider extending eNVQM to take into account user profile, when it is available, studying the effect of employing congestion control mechanisms and application layer adaptive solutions for delivering 3D video content and performing additional tests involving the latest HEVC/H.265 standard. It will also conduct additional subjective tests using one and both views of the 3D video sequences subject to packet loss in order to determine at what level of loss it is better to switch from 3D to 2D viewing. Finally, the impact of video content type on the results of the no-reference quality assessment will also be studied.

ACKNOWLEDGMENTS

This research was supported in part by the Telecommunications Graduate Initiative (TGI), funded by the Higher Education Authority under the Programme for Research in Third-Level Institutions (PRTLI) Cycle 5, and co-funded under the European Regional Development Fund (ERDF), by the Science Foundation Ireland grant no. 10/CE/I1855 to LERO, the Irish Software Engineering Research Centre and by Electronic Science and Technology-Zhejiang Open Foundation of the Most Important Subjects under GK130203207003/006.

REFERENCES

[1] H.265, "High efficiency video coding", Apr. 2015.

[2] International 3D & Advanced Imaging Society. (2015). "3D@Home Website – Steering Teams Overview" [Online]. Available: http://www.3dathome.org/steering-overview.aspx.

[3] China 3D Industry Association, "China 3D Industry Association" Aug. 2015. [Online]. Available: http://www.c3dworld.org/.

[4] 3D-Consortium, "3D Consortium-New era of 3D- 'From surprise to impression!' " Aug. 2014. [Online]. Available: http://www.3dc.gr.jp.

[5] Tam, W. J., Speranza, F., Yano, S., Shimono, K. and Ono, H. "Stereoscopic 3D-TV: Visual Comfort," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp.335-346, Jun. 2011.

[6] Lin, Y. H. and Wu, J. L. "A depth information based fast mode decision algorithm for color plus depth-map 3D videos," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2), 542-550, 2011.

[7] Pulipaka, A., Seeling, P., Reisslein, M., Karam, L.J., "Traffic and Statistical Multiplexing Characterization of 3-D Video Representation Formats," *Broadcasting, IEEE Transactions on*, vol. 59, no. 2, pp.382-389, Jun. 2013.

[8] Merkle P., Müller K., and Wiegand T., "3D Video: Acquisition, Coding, and Display," *Proc. IEEE International Conference on Consumer Electronics (ICCE '10)*, Las Vegas, USA, Jan. 2010.

[9] Park, M., Luo, J., Gallagher, A. C., and Rabbani, M., "Learning to Produce 3D Media From a Captured 2D Video," *Multimedia, IEEE Transactions on*, vol.15, no.7, pp. 1569-1578, Nov. 2013.

[10] Gopalakrishnan, V., Bhattacharjee, B., Ramakrishnan, K.K., Jana, R., Srivastava, D., "CPM: Adaptive Video-on-Demand with Cooperative Peer Assists and Multicast," *IEEE INFOCOM*, pp. 91-99, Apr. 2009.

[11] Yuan, Z. and Muntean, G.-M., "A Prioritized Adaptive Scheme for Multimedia Services over IEEE 802.11 WLANs," *IEEE Transactions on Network and Service Management*, no.4, vol.10, pp. 340 - 355, Dec. 2013.

[12] Yuan, Z., Venkataraman, H. and Muntean, G.-M. , "iPAS: An User Perceived Quality-based Intelligent Prioritized Adaptive Scheme for IPTV in Wireless Home Networks," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Shanghai, China, pp. 1-6, Mar. 2010.

[13] Xu, C., Jia, S., Zhong, L., Zhang, H. and Muntean, G. M. "Ant-Inspired Mini-Community-Based Solution for Video-On-Demand Services in Wireless Mobile Networks," *Broadcasting, IEEE Transactions on*, vol. 60, no. 2, pp.322-335, Jun. 2014.

[14] ITU-T G.1070, "Opinion model for video-telephony applications," Jul. 2007.

[15] Yasakethu, S. L. P., Hewage, C. T., Fernando, W. A. C., and Kondoz, A. M., "Quality analysis for 3D video using 2D video quality models," *IEEE Trans. Consum. Electron.* vol. 54, no. 4, pp. 1969-1976, Nov. 2008.

[16] Hewage, C. T., Worrall, S. T., Dogan, S., Villette, S., and Kondoz, A. M., "Quality evaluation of colour plus depth map based stereoscopic video," IEEE. J. Sel. Topics Signal Process. vol. 3, no. 2, pp. 304-318, Apr. 2009.

[17] Lu, F., Wang, H., Ji, X., and Er, G., "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model," in *Proc. 3DTV Conference*, Potsdam, Germany. pp. 1-4, May 2009.

[18] Ozbek, N., and Tekalp, A. M., "Unequal inter-view rate allocation using scalable stereo video coding and an objective stereo video quality measure," in *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 1113-1116, Monterrey, México, Apr. 2008.

[19] Shao, H., Cao, X., and Er, G., "Objective quality of depth image based rendering in 3DTV system," *Proc. 3DTV Conference*, pp. 1-4, May 2009.

[20] Joveluro, P., Malekmohamadi, H., Fernando, W. A C, Kondoz, A.M., "Perceptual Video Quality Metric for 3D video quality assessment," *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1-4, Jun. 2010.

[21] Huynh-Thu, Q., Ghanbari, M., "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol.44, no.13, pp. 800-801, Jun. 2008.

[22] Wang, Z., Lu, L., and Bovik, A. C., "Video quality assessment using structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121 -132, 2004.

[23] Pinson, M. H., and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312 -322, Sept. 2004.

[24] Shao, F., Jiang, G., Lin, W., Yu, M., Dai, Q., "Joint Bit Allocation and Rate Control for Coding Multi-View Video Plus Depth Based 3D Video," *Multimedia, IEEE Transactions on*, vol.15, no.8, pp. 1843-1854, Dec. 2013.

[25] Lee S.-B., Muntean, G.-M., Smeaton, A.F., "Performance-aware replication of distributed pre-recorded IPTV content," *IEEE Transactions on Broadcasting*, vol.55, no.2, pp.516-526, Jun. 2009.

[26] Han Y., Yuan Z., Muntean, G.-M., "No reference objective quality metric for stereoscopic 3D video," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1-6, 25-27, Jun. 2014.

[27] Han, Y., Yuan, Z., Muntean, G.-M., "Extended No Reference Objective Quality Metric for Stereoscopic 3D Video," IEEE ICC on Quality of Experience-based Management for Future Internet Applications and Services, London, UK, Jun. 2015.

[28] Kooi, F. L., and Toet, A. "Visual comfort of binocular and 3D displays," *Displays*, 25(2), pp. 99-108, 2004.

[29] ITU-R BT.1198, "Stereoscopic television based on R-and L-eye two channel signals", Oct. 1995.

[30] ITU-T BT.500, "Methodology for the subjective assessment of the quality of television pictures", Jan. 2012.

[31] ITU-T R. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment", Apr. 2014.

[32] ITU-T P.3D-sam, "Subjective assessment methods for 3D video quality", Jul. 2015.

[33] ITU-T J.3D-fatigue, "Assessment methods of visual fatigue and safety guideline for 3D video", 2015.

[34] ITU-T J.3D-disp-req, "Display requirements for 3D video quality assessment", 2015.

[35] Hewage, C.T., Martini, M.G., "Quality of experience for 3D video streaming," *Communications Magazine, IEEE* , vol.51, no.5, pp. 101-107, May 2013.

[36] Wang, K., et al., "Stereoscopic 3D video coding quality evaluation with 2D objective metrics," *In Proc. SPIE Electronic Imaging*, vol. 8648, Mar. 2013.

[37] Xing, L., You, J., Ebrahimi, T., Perkis, A., "Assessment of Stereoscopic Crosstalk Perception," *Multimedia, IEEE Transactions on*, vol.14, no.2, pp. 326-337, Apr. 2012.

[38] Feitor, B., Assuncao, P., Soares, J., Cruz, L., Marinheiro, R., "Objective quality prediction model for lost frames in 3D video over TS," *IEEE International Conference on, Communications Workshops (ICC)*, pp. 622-625, Jun. 2013.

[39] Soares, J.R.S., da Silva Cruz, L.A., Assuncao, P., Marinheiro, R., "No-reference lightweight estimation of 3D video objective quality," *IEEE International Conference on, Image Processing (ICIP)*, pp. 763-767, 27-30 Oct. 2014.

[40] ITU-T P.800, "Methods for subjective determination of transmission quality", Jun. 1998.

[41] Kawano, T., Yamagishi K., and Hayashi T.. "Performance comparison of subjective assessment methods for 3D video quality." *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on. IEEE*, 2012.

[42] Adler, J., & Parmryd, I. "Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient". *Cytometry Part A*, 77(8), pp.733-742. 2010.

[43] ITU-T R. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," Apr. 2014.

[44] Carbone M. and Rizzo L., "Dummynet Revisited", *ACM SIGCOMM Computer Communication Review*, 40(2) , p.12-20, Mar.2010.

[45] Sarkar, B., & Menq, C. H. "Smooth-surface approximation and reverse engineering", *Computer-Aided Design*, 23(9), 623-628, 1991.

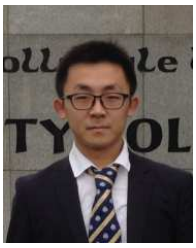
[46] MSU Video Quality Measurement Tool. [Online]. Available at: http://compression.ru/video/quality_measure/video_measurement_tool_en.htm.

[47] Zinner, T., Abboud, O., Hohlfeld, O., Hossfeld, T., and Tran-Gia, P. (2010, March). "Towards qoe management for scalable video streaming," *21th ITC Specialist Seminar on Multimedia Applications-Traffic, Performance and QoE* , pp. 64-69, Mar. 2010.

[48] Dymarski, P., Kula, S., and Huy, T. N. (2011). "QoS Conditions for VoIP and VoD," *Journal of Telecommunications & Information Technology*, vol. 2011, no.3, pp.29-37, Mar. 2011.

[49] IJsselsteijn, W., de Ridder, H., Hamberg, R., Bouwhuis, D., and Freeman, J. "Perceived depth and the feeling of presence in 3DTV," *Displays*, 18(4), 207-214, 1998.

[50] Benoit, A. Le Callet, P. Campisi P., and Cousseau R., "Quality Assessment of Stereoscopic Images," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–13, 2008.



Yi Han received his Bachelor of Engineering degree at International School of Software, Wuhan University, China in 2010, and Master in Telecommunication at Dublin City University in Sep. 2011. Since then he is pursuing his PhD with the Performance Engineering Laboratory, located in both University College Dublin and Dublin City University, Ireland. He performs research in the area of quality assessment and adjustment of 3D video and Voice/Video over IP deliveries.

Dr. Zhenhui Yuan (S'09-M'12) is with the Ministry of Education Key Frequency and Circuit Lab, School of Electronic and Information, Hangzhou Dianzi University, China. He was a Postdoctoral Researcher with the Performance Engineering Laboratory, School of Electronic Engineering, Dublin City University, Ireland, which has also awarded him the Ph.D degree for research in video delivery differentiation in wireless network environments in 2012. In 2008, Zhenhui received his B.Eng from Wuhan University China. His research interests include wireless mobile networks, quality-oriented and battery-aware multimedia streaming, and multisensorial services. He is a member of IEEE.

Dr. Gabriel-Miro Muntean (S'99-M'04) is a Senior Lecturer with the School of Electronic Engineering, co-Director of the Performance Engineering Laboratory at Dublin City University, Ireland and Consultant Professor with Beijing University of Posts and Telecommunications, China. He obtained his Ph.D. degree from Dublin City University, Ireland for research in quality-oriented adaptive multimedia streaming over wired networks in 2003. He was awarded the B.Eng. and M.Sc. degrees in Software Engineering from the Computer Science Department, "Politehnica" University of Timisoara, Romania in 1996 and 1997, respectively. Dr. Muntean's research interests include quality and performance-related issues of adaptive multimedia streaming, performance of content delivery over wired and wireless networks and with various devices, and energy-aware networking. Dr. Muntean has published over 250 papers in top-level international conferences and journals and has authored three books and sixteen book chapters and edited six other books. Dr. Muntean is Associate Editor with the IEEE Transactions on Broadcasting, Associate Editor with the IEEE Communications Survey and Tutorials and reviewer for important international journals, conferences and funding agencies. He is a member of ACM, IEEE and IEEE Broadcast Technology Society.