

3D video subjective quality: a new database and grade comparison study

Emil Dumić¹ · Sonja Grgić² · Krešimir Šakić³ ·
Pedro Miguel Regalo Rocha⁴ · Luis A. da Silva Cruz⁴

Received: 22 July 2015 / Revised: 30 October 2015 / Accepted: 15 December 2015
© Springer Science+Business Media New York 2016

Abstract This paper presents a research study on the subjective assessment of 3D video quality using a newly constructed 3D video database (3DVCL@FER). This database consists of 8 original 3D video sequences, each degraded with 22 different degradation types, including degradations specific to stereoscopic systems. The subjective assessment was done with the support of a purpose-built easily customizable grade collection platform and conducted in two research laboratories, in Croatia and Portugal. Subjective scores for quality, depth and comfort were collected and DMOS (Difference Mean Opinion Score) values were calculated. Different objective measures (for image, 3D image, 2D video and 3D video) were separately compared with DMOS values for quality, depth and comfort. The 3D video grade-annotated database described is publicly accessible and can be used in research-related activities like assessment of existing objective measures, using the entire database or parts of it, and construction of new objective measures specific to 3D video degradations. The system presented can also be used to collect and compare subjective quality grades originating from different sites to study the effect of different observation conditions and observer/graders populations on the DMOS quality values for 3D video depth and comfort.

Keywords 3D video quality · Subjective assessment · Objective quality measures · 3DVCL@FER video database

✉ Emil Dumić
emil.dumic@gmail.com

¹ Department of Electrical Engineering, University North, 104. brigade 3, 42000 Varaždin, Croatia

² Faculty of Electrical Engineering and Computing, Department of Wireless Communications, University of Zagreb, Unska 3/XII, HR-10000 Zagreb, Croatia

³ Radio Communication Department, Croatian Regulatory Authority for Network Industries (HAKOM), Zagreb, Croatia 10000

⁴ Department of Electrical and Computer Engineering, University of Coimbra, Rua Sílvio Lima, Pólo II, Coimbra, Portugal 3030-790

1 Introduction

Subjective quality assessment of video sequences deals with how the quality of video sequence is perceived by an observer. In a typical subjective quality assessment campaign, test subjects watch a number of original and degraded video sequences and rate their quality on a numeric scale. Subjective quality is often expressed as Mean Opinion Score (MOS) that represents a quality grade attributed by a standard average observer to a given video sequence. MOS grades are collected following well defined methods and procedures that have been proposed in the last decades and which aim at guaranteeing that the same experimental settings and conditions are used during different assessment campaigns. One of the most well-known and used subjective (2D) video quality assessment methods is described in ITU-R BT.500-13 [25]. This ITU recommendation defines several types of test methods, which can be generally divided in double stimulus or single stimulus methods, depending on the method how the content is shown to the observer. The related ITU-R BT.2021 [24] recommendation is based on the methods described in ITU-R BT.500-13 [25] expanded and adapted for subjective assessments of stereoscopic 3DTV systems. The main difference between subjective assessment of 2D and 3D video sequences is that in 3D, besides the general video quality grade, two additional grades measuring depth quality and visual comfort are asked from the test subjects. The other notable difference is that for 3D video quality assessment, stereoscopic vision of the test subjects needs to be verified alongside color and vision acuity. These additional verifications are important because subjects' visual fitness has a significant impact on the accuracy and precision of the subjective quality grades.

In this article we describe a web-based system built to facilitate the collection of 3D video subjective quality grades at remote locations, which follows some of the procedures defined in the latter of the aforementioned recommendations. Another important contribution of this work is a freely accessible ensemble of (possibly degraded) 3D video sequences annotated with their respective quality grades, collected using this system deployed at two research laboratories in two countries. This database will henceforth be designated by 3DVCL@FER and can be accessed electronically as described later. An analysis of the grades compiled and included in the database is also presented together with some comments and conclusions. Evaluation of different objective measures (for image, 3D image, 2D video and 3D video) was also done, using DMOS scores for quality, depth and comfort separately.

The text is organized as follows. Section 2 summarizes related work regarding subjective and objective assessment of 3D video content. Section 3 gives an overview of the 3DVCL@FER video database describing its constituent video sequences. Section 4 presents the web-based application, which was designed and used for this subjective assessment grade collection project. Section 5 lists the results of the subjective assessment, while section 6 analyses the results and discusses the use of the 3DVCL@FER database in similar works focused on the estimation of 3D video quality. Finally, section 7 presents our conclusions.

2 Related work

Objective video quality assessment methods are often used during the course of designing a video communication system and in other applications where there is a constant need for assessment of the effect on final video quality of various algorithmic optimizations, content variations and transmission conditions. In general objective quality measures can be divided

into three categories according to the reference information they use: Full reference (FR) measures that require the original undistorted or unprocessed signal, reduced-reference (RR) measures which make use of information derived from the original signal and no-reference (NR) quality measures that rely only on the processed/degraded signal. It has been observed that traditional objective quality measures [10] such as Peak Signal to Noise Ratio (PSNR) or Structural Similarity (SSIM) [39] do not correlate well with subjective grades for degraded 3D video sequences [9]. To address this problem new objective methods [1, 3, 27], are being developed which are better suited for 3D video quality assessment and take into account the supplemental depth information specific to 3D video. These research activities create a need for new 3D video sequence databases annotated with the corresponding subjective quality grades that can then be used to assess the performance of current and new objective quality prediction methods. To compare the proposed objective measures, usually MOS scores are processed to obtain DMOS (Difference MOS) grades, which quantify the subjective quality difference between original and degraded sequences (as FR and RR objective measures do). Earlier methods proposed for assessing 3D video and image quality drew directly from 2D image and video quality measures with some adaptations to cope with the specificity of the representation formats, e.g. in the case of 3D stereo video by applying a 2D measure to both left and right views and combining the results. The performance of these adaptations depends heavily on the type of content and degradations involved. For instance in [31], the authors concluded that 2D image quality measures applied to a 3D image database with symmetrically induced distortions performed better than 3D image quality measures. However, in [4], a cyclopean model was developed and added as an extension to 2D image quality measures, which then outperformed basic 2D measures, when tested on asymmetrically induced distortions of 3D distorted images. These experiments as well as several others strongly suggest that particular care should be taken, when comparing objective and subjective measures, because 3D measures will not always give better correlation with subjective grades than 2D measures, especially if they were not developed for the specific degradations affecting the 3D content to be graded.

In [30] the authors proposed a novel distortion metric specifically designed for use in rate-distortion optimized encoding of the depth information of 3D video contents. It achieves competitive performance with state of the art rate distortion optimization algorithms, while maintaining a relatively lower complexity. In [29] the authors have proposed techniques to estimate the objective quality of stereoscopic 3D video content, which have as inputs the motion and depth map features of 3D video. They concluded that information provided by depth maps and estimated motion vectors is useful to model the visual discomfort and fatigue experienced by observers when contemplating 3D stereoscopic contents. In [2] authors studied the relationship between 3D quality and bitrate at different frame rates. In first part, they concluded that higher frame rates (48 or 60 fps) are preferred over lower frame rates (24 or 30 fps). In second part, they concluded that higher frame rates with higher compression rates are preferred over low frame rate with slightly compressed 3D videos, for 3D video sequences with fast motion.

Various image and video databases have been developed in recent years due to the different needs of different research works on subjective and objective assessment of multimedia quality. Most of the databases that can be found in the literature or on the Internet are described in [34]. An analysis of this compilation and of the description of the databases itemized shows that most 3D video databases listed have been developed for a specific purpose only, in most cases including grades only for the quality of the content, omitting the depth and comfort

grades specified by ITU-R BT.2021. That study also shows that 3D image and video databases are for now less researched and in lesser numbers than their 2D counterparts. Some examples of 3D stereoscopic video datasets are those from [5, 13, 32, 37, 41]. To the author's best knowledge, none of the existing 3D video databases has been designed to include a large number of different distortion types (both 2D and 3D specific), together with quality, depth and visual comfort grades.

In the next sections we describe the main contributions of this work, starting with a description in Section 3 of a 3D video subjective quality grade database (3DVCL@FER). As will be explained in Section 4, the way the system is designed and used can help mitigate the problems listed before, encountered when collecting large sets of subjective grades using pure laboratory based approaches or pure web-based solutions.

3 The formation of 3D video database

To form new 3D video database we selected 8 original 3D stereo video sequences that are available for download from [12] (src01-src08) as well as derived sequences with specific degradations: Car and barrier gate, Basketball training, Boxers, Hall, Laboratory, News report, Phone call, Soccer. All the 8 original sequences are in full HD stereo format, with 25 fps frame rate and are 16 seconds long. First frame from each sequence, left view, is presented in Fig. 1.

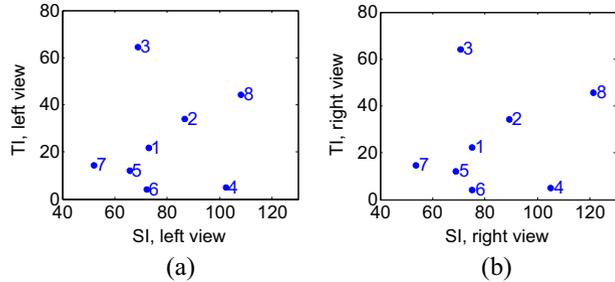
The dynamic characteristics of the reference sequences, as measured by spatial and temporal activity indices were computed on the left and right view according to the procedure defined in ITU-T recommendation P.910 [26]. The sequences activity indices are plotted in Fig. 2a) (left view) and Fig. 2b) (right view) and are labeled from 1 to 8. The plots show that the sequences are very diverse in terms of their dynamic characteristics.

These eight original sequences included in the database were complemented with several sequences showing the effect of specific types of degradation. A total of 22 degradation types, listed in Table 1, were applied to each of the eight original sequences, so that besides the original eight sequences, our 3D video database contains 176 sequences with degradations. We called our new 3D video database "3DVCL@FER". Depending on their type, the degradations were generated either using ffmpeg-x64 (version 22.11.2014) downloaded from [15] or the H.264/AVC encoder from the JVT JM18.6 reference software package with settings derived from those specified in [38]. Degradation types 21 and 22, based on a 3D-HEVC



Fig. 1 First frame, left view, from each of the tested sequences

Fig. 2 Spatial versus temporal information: (a) left view; (b) right view



encoding, differ from degradation types 11 and 12, based on H.264/AVC reference encoder, because 3D-HEVC encodes left view as a base view and the right view as a dependent view, using HEVC core encoding tools for both intra-view and inter-view. H.264/AVC as used in this work encodes independently the left and right views.

The grade information lists 146 grades collected in a laboratory of Faculty of Electrical Engineering and Computing in Croatia. Additional grades were collected in a laboratory in Portugal, but due to its smaller numbers for the time being they were not included in the study. The data provided in the database comprises the sequences in uncompressed format (.avi separate left+right) and in a near-losslessly compressed format (x.264+.mp4 and vp8+.webm combined left+right). The entire 3DVCL@FER database including all the sequences and grades described above is publicly available and can be downloaded from the location at [16].

4 Subjective study set up using web-based application

A web-based application was developed for the purpose of conducting the subjective assessments of the 3DVCL@FER database contents. The application is easily customizable and can be used with different web browsers. In the case-study reported here it was setup to be used with Google Chrome and Mozilla Firefox web browsers. It was programmed using the javascript and php languages and customized to display 3D video on computers equipped with a 3D monitor. The application collects and saves the subjective scores in a results database.

The start page of the application (shown in facsimile in Fig. 3) presents instructions about the testing procedure to the test subject. This GUI is also used to collect some information about the test subject such as age, gender and e-mail address and some additional information about the monitor type.

Several control mechanisms are implemented in the application to ensure the validity of the scores collected. The most important one is that the application switches automatically to full screen during the whole duration of the assessment. If the test subjects exit the full screen mode the test procedure automatically stops and the corresponding result is discarded. Only the results from the subjective assessment that run from start to finish in full screen mode are flagged as valid, stored in the result database and used in the final results analysis.

Although the results were collected in laboratory conditions in Croatia and Portugal, application can be also used for pure web-based evaluation, which is planned in the future. Application is developed to be use in two different setups: one is by using Chrome web browser and 3DTV with capability of manual switching to 3D mode [6], and the other is by

Table 1 Degradation type in 3DVCL@FER database

Number	Degradation	Detailed description	Tools and settings
1	2D view	left+right view becomes left+left view	–
2	resizing	4x4 down and up using lanczos3 filter	ffmpeg-x64
3	frame rate reduction	to 1/3 of the original fps	ffmpeg-x64
4	brighten	y value+15, right view only	ffmpeg-x64
5	change gamma	to 0.6, right view only using	ffmpeg-x64
6	horizontal disparity	left view 30 pixels left, right view 30 pixels right	ffmpeg-x64
7	horizontal disparity	left view 30 pixels right, right view 30 pixels left using	ffmpeg-x64
8	vertical disparity	left view 20 pixels down, right view 20 pixels up	ffmpeg-x64
9	geometric distortion	left view only	stirmark [20]
10	2D to 3D conversion	2D to 3D conversion	ffmpeg+avisynth downloaded from [19]; avisynth script based on [42]
11	H.264/AVC coding	QP = 32 (both views): specific setting QP = 32	H.264/AVC reference encoder and decoder version 18.6, downloaded from [18]
12	H.264/AVC coding	QP = 44 (both views): specific setting QP = 44	H.264/AVC reference encoder and decoder version 18.6
13	H.264/AVC coding	left QP = 32, right QP = 44: specific setting for QP parameter (asymmetric)	H.264/AVC reference encoder and decoder version 18.6
14	H.264/AVC coding	QP = 32 with edge enhancement, strength 75 %	avisynth and toon filter; toon filter can be downloaded from [14]; H.264/AVC reference encoder version 18.6; ffmpeg-x32 version 12.6.2014 downloaded from [15] (x264 decoder only)
15	packet losses	generated with Gilbert-Elliott model; specific parameters for Gilbert-Elliott model: $\alpha = 0.01$, $\beta = 0.1$, $\gamma = 0.4$, $\delta = 0.01$	H.264/AVC reference encoder version 18.6; sirannon software downloaded from [23]; sirannon script proposed in [38]; error concealment and x264 decoder using ffmpeg-x64 (-ec switch set to 2)
16	2D view, H.264/AVC coding	2D left view only, QP = 44	H.264/AVC reference encoder and decoder version 18.6 downloaded from [18]
17	jpeg2000 compression	bitrate 2 Mbps	JPEG2000 kakadu software downloaded from [21]
18	frame-freeze	2 s long, online streaming (degraded video is same duration as original)	ffmpeg-x64 and avisynth
19	frame-freeze	2 s long, offline streaming (degraded video is longer than original)	ffmpeg-x64 and avisynth
20	3D to 2D switching	(left+right becomes left+left view) back to 3D, 2 s long, degraded video is same duration as original	ffmpeg-x64 and avisynth

Table 1 (continued)

Number	Degradation	Detailed description	Tools and settings
21	3D-HEVC encoding	QP = 32: General configuration settings are based on “baseCfg_2view.cfg” in the same software, with specific setting for QP factors	HTM reference encoder and decoder 3D-HEVC version 11.0, downloaded from [17]
22	3D-HEVC encoding	QP = 44: General configuration settings are based on “baseCfg_2view.cfg” in the same software, with specific setting for QP factors	HTM reference encoder and decoder 3D-HEVC version 11.0, downloaded from [17]

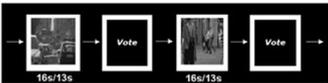
using Firefox browser and 3D monitor with an Nvidia 3D vision system [7]. Also, provided links use preload option, which means that all 3D content is firstly stored on the local machine, prior running the test. This was done by using Chrome or Firefox cache and it is very useful because this makes evaluation independent on the download speed (of the machine on which the test is being done). However, this makes overall duration of the test at least twice as long, compared to the case in which preloading would not be used. Details of both setups can be found on the starting web pages [6] and [7]. Preloading option could be also skipped, which is useful if 3D content is already stored on the machine from which the test is being done (e.g. in laboratory), or if server and client machine are both on high speed network (preferably 1 Gbit/s). We have also skipped preloading in both laboratories.

The assessment of the subjective quality of the 3D videos from the 3DVCL@FER database performed using the system is based on Absolute Category Rating (ACR) with hidden reference (ACR-HR). In ACR-HR, each original unimpaired signal is included in the experiment but not identified as such. The ratings for the original signals are removed from the



The test will last about 10-15 minutes during which you will grade 28 differently degraded and original 3D video sequences in random order. You can exit the test at any time by pressing “Esc”. Afterwards, the page will be reloaded. However, these grades won't be used in any further experiment.

The test procedure is shown on the image below. After every sequence, you will be asked to grade its picture quality, depth quality and visual comfort.



You have to **grade every** video sequence with the help of the slider shown on the right image. The slider will appear after the end of every video sequence. It has a span between 0 and 5 with the step of 0.1.

For picture quality and depth quality grade 0 represents bad, while 5 represents excellent.
 For visual comfort grade 0 represents extremely uncomfortable while 5 represents very comfortable.



After you choose your desired grade, push the button GRADE. You can use your mouse and/or “TAB” and “SPACE” key to navigate between sliders and “arrows” or mouse to move the slider. Next video sequence will start after you click the button GRADE, using mouse or “ENTER” key. Every grade has to describe your **own** opinion about the picture quality, depth quality and visual comfort of the tested video sequence. Therefore, grade its quality over the overall duration of the video. You can give very low or very high grades if that represents your opinion.

Also, please take the test without any other distracting activities. Thank you once again for your cooperation.

Please enter the following information (not necessary)

E-mail:	<input type="text"/>	Enter your email address.
3D monitor type:	<input type="text"/>	Enter your 3D monitor type (if known).
Illumination type:	<input type="text"/>	sunlight, fluorescent, incandescent, LED.
Time of the day:	<input type="text"/>	day or night.
Age:	<input type="text"/>	Enter your age (numbers).
Gender:	<input type="text"/>	Please write male or female.

Switch to 3D mode, push the “BEGIN” button when you are ready and when preload is finished:

BEGIN

Fig. 3 Start page of the application used for 3D subjective quality assessment

scores of the associated processed video sequences during data processing [26]. The grading is done on three different dimensions, each one graded on a continuous scale from 0 to 5 with a step of 0.1 according to the [24]. The three dimensions represent picture quality, depth quality and visual comfort. For picture quality and depth quality grade 0 represents bad, while 5 represents excellent. For visual comfort grade 0 represents extremely uncomfortable while 5 represents very comfortable.

In order to reduce the time needed to conduct the subjective assessment to comfortable durations, the 184 video sequences were divided into 8 sets. Each set contains 4 original and 22 associated degraded sequences (26 sequences per test). At the beginning of the test, we added 2 additional sequences intended to serve as an introduction to the observers of the grading system and assessment procedure and to serve as a qualification test. The results associated with those 2 sequences are removed in later processing of the results. In this test procedure, one subjective assessment session takes $28 \times 16 = 448$ s plus the time needed for the observer to grade the just-observed sequences. Every segment includes all 22 degradations, so no further realignment would be necessary (due to the different degradations per session), when calculating MOS or DMOS scores. Indeed since during each session all types of degradations are observed only once, the human visual system (HVS) does not get tuned to any specific distortion type.

The tests were conducted in two research laboratories in Croatia and Portugal, with somewhat different setups and two slightly different versions of the web application. The differences are mostly related to support of different monitor types and 3D setups. Slightly different video sequences sets were chosen in accordance with the different setups used in the two laboratories to cope with different capabilities of the monitors and internet browsers used. Prior to the subjective evaluation, all graders involved in the study in Portugal were screened for visual acuity, color vision and stereo vision using Snellen chart, Ishihara chart and “Fly” test, respectively. In Croatia, all subjects were screened for color vision and stereo vision using Ishihara chart and Randot test.

The 184 sequences used in Croatia, were compressed at a high quality setting, using the x264 encoder (in .mp4 container, left+right view) with constant rate factor (CRF) 10, to make them playable in the Chrome browser. Additionally, in order to validate the test setup and verify that the compression used to permit running the test over the internet did not negatively impact the quality scores, PSNR and SSIM were calculated with the uncompressed sequence as reference. The following results were observed: minimum PSNR=47.004 dB, median PSNR=50.9415 dB, mean PSNR=51.5362 dB, maximum PSNR=58.6240 dB; minimum SSIM=0.9920, median SSIM=0.9959, mean SSIM=0.9959, maximum SSIM=0.9996. These results show clearly that the H.264/AVC compressed video sequences have near-lossless quality and so the compression used will not bias the scores collected from the evaluation sessions.

The monitor type used in the subjective assessments in Croatia is a 32 inch Samsung UE32H6400 with active shutter glasses and with manual switch to 3D capability. The monitor settings used were the factory defaults for standard 3D, left+right input video format. The computer used in the work was equipped an Intel i7-4790 k processor, Solid State Drive (SSD), 16 GB of RAM and running Windows 8.1. An example test can be run from [6]. The configuration accessible through the link uses a preloading option, which is useful for lower download speeds or web-based test.

The grading sessions that took place in the Croatian laboratory were done with fluorescent illumination (artificial lighting) under controlled constant conditions and 6500 K color

temperature. The observers were mostly students, both male and female, 20–44 years old (24.38 on average). Average distance from the display was about 1.2 m.

The 184 video sequences used in Portugal were encoded with the vp8 encoder with CRF 6, to make them playable in the Firefox browser with an Nvidia 3D vision system. The PSNR values calculated between the original and compressed sequences gave the following results: minimum PSNR=45.0800 dB, median PSNR=47.9600 dB, mean PSNR=48.4989 dB, maximum PSNR=54.6800 dB. It can be concluded that the vp8 compressed video sequences are near-lossless in terms of quality and so can be used without fear of biasing the results.

The 3D video monitor used in Portugal is a 27 inch Asus VG278 HR with active shutter glasses. The monitor settings used were the factory defaults for standard 3D. The computer used in the work was equipped an Intel Xeon E5-1607 @ 3.00 GHz processor, 16 GB of RAM and running Windows 7. An example of the test can be run from [7], which also uses the preloading option.

Tests in Portugal laboratory were performed with fluorescent illumination (artificial lighting) under controlled, constant conditions. Observers were mainly students, both male and female, 21–48 years old (24.66 on average).

It should be noted that at the time of writing only 35 observers' scores had been collected in the Portuguese laboratory. Also, we had only 2–6 grades per video sequence ($35/8=4.4$ grades on average). This means that inter-laboratory DMOS correlation studies as well as correlation with different objective measures could be biased. For this reason, later comparisons with subjective and objective measures used only results collected in the Croatian laboratory, therefore ensuring that enough grades per sequence were collected, according to the recommendations of ITU-R BT.500-13 [25].

5 Results of subjective testing

After the conclusion of the grading sessions, the scores collected from Croatian laboratory were converted to DMOS and compared with seven objective measures. As it is not yet very well known how to interpret and process raw scores for depth and comfort to calculate DMOS values, those scores were treated as usual quality scores. Overall we gathered 146 observations, resulting on an average of $146/8\approx 18$ grades per video sequence, before elimination of outliers. The MOS scores collected together with confidence intervals (according to Fisher's least significant difference procedure) are presented in Fig. 4, before outlier removal. Histogram of MOS scores for quality (also before outlier removal), divided in 7 equal bins, is presented in Fig. 5.

We applied chi-square goodness of fit for uniform distribution (for MOS scores for quality), divided in 7 equal bins. Chi-square goodness of fit tests against null hypothesis that the observed data (in our case MOS quality scores) are a random sample from a tested distribution (in our case uniform) with mean and variance estimated from the data, against the alternative that the data are not from the tested distribution with the estimated mean and variance. We obtained p value 0.2137 which means that null hypothesis cannot be rejected and that MOS quality scores have uniform distribution across the whole range (0.47–4.17), and so can be used in further experiments (e.g. in comparison with different objective measures). However, it should be noted that p value depends also on the number of bins in which we divide range. With higher number of bins (10 and more), p value would be lower than 0.05, resulting in null hypothesis rejection.

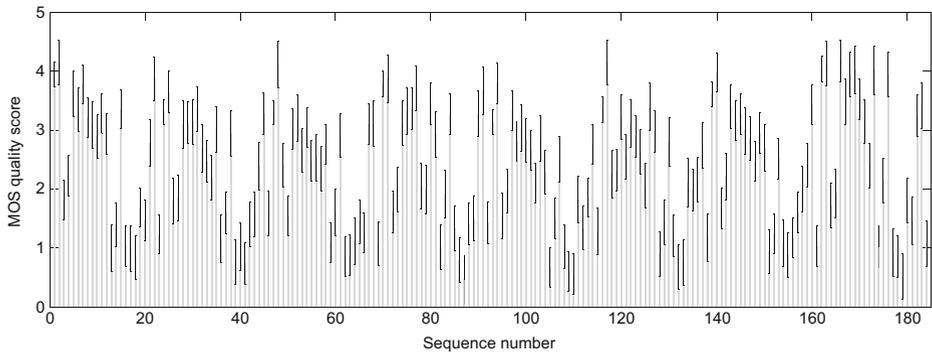


Fig. 4 MOS quality scores

Afterwards, we converted observer's grades to DMOS scores and removed outliers. We applied the same procedure to the observer's scores for quality, depth and comfort. DMOS scores are useful when comparing with full reference or reduced reference objective measures (which we will use later). For no reference objective measures, MOS scores could provide better correlation, so those scores can be calculated from the data in [16].

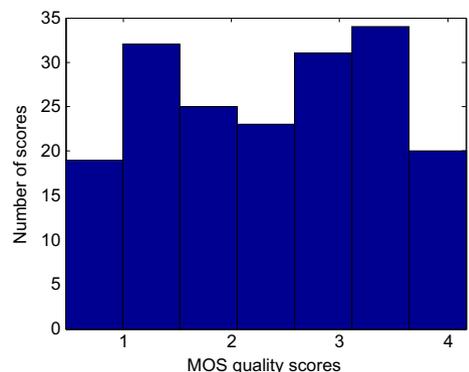
The first 2 sequences (qualification test) were ignored and not used in these calculations. Screening of the observers was performed according to the ITU-R BT.500-13 [25] to discard scores from observers who differ too much from the average value. Each residual (difference between reference and degraded video sequence's grade from the same observer) was converted to z-score according to (1).

$$z_{nl} = \frac{d_{nl} - \mu_n}{\sigma_n} \quad (1)$$

In (1) z_{nl} is the z-score of observer n , for video sequence l , d_{nl} is the residual score of observer n , for video sequence l , μ_n is the mean score from observer n and σ_n is standard deviation of scores from observer n (over all tested sequences l for that observer). This normalization is done to remove the effects of any differences in the use of the quality scale (differences in the location and range of values used by the observer).

For each time window (16 s per video sequence) normality of the z-scores was tested using kurtosis β , over the span of all z-scores from the particular video sequence. Depending on the

Fig. 5 Histogram of scores for quality, divided in 7 equal bins



kurtosis, each observer's grade was compared to a multiple of the deviation σ_l from the mean value of each video sequence l . Following recommendation ITU-R BT.500-13 [25], the decision of whether or not the data from a given observer should be discarded is based on two values, P and Q , computed according to (2).

$$\left. \begin{array}{l} \forall l \in L \text{ where } L \text{ stands for number of video sequences} \\ \forall n \in N \text{ where } N \text{ stands for number of observers} \\ \text{if } z_{nl} \geq \bar{z}_l + 2 \cdot \sigma_l \text{ then } P_n = P_n + 1 \\ \text{if } z_{nl} \leq \bar{z}_l - 2 \cdot \sigma_l \text{ then } Q_n = Q_n + 1 \end{array} \right\} \text{for } 2 \leq \beta \leq 4 \text{ (normal)} \quad (2)$$

$$\left. \begin{array}{l} \text{if } z_{nl} \geq \bar{z}_l + \sqrt{20} \cdot \sigma_l \text{ then } P_n = P_n + 1 \\ \text{if } z_{nl} \leq \bar{z}_l - \sqrt{20} \cdot \sigma_l \text{ then } Q_n = Q_n + 1 \end{array} \right\} \text{for } \beta \notin [2, 4] \text{ (not normal)}$$

These P and Q values are computed for every observer and if any of them is larger than 10 % of the number of tested degraded video sequences (in our case 22/10, which is rounded to 3 sequences per observer), that observer's data is discarded. Using this method, 3 observers were removed from further analysis for quality MOS, 2 observers for depth MOS and 0 observer for comfort MOS, effectively removing a total of 5 observers from the pool.

Afterwards, results for every observer were rescaled to the full range of 0–100, according to (3) where $\max(z)$ and $\min(z)$ represent maximum and minimum z -scores over all observers and all video sequences):

$$dmos_{n,l} = \frac{100}{\max(z) - \min(z)} \cdot (z_{n,l} - \min(z)) \quad (3)$$

and $dmos_{n,l}$ represents the rescaled grade of viewer n and sequence l . At the end, an average DMOS grade was calculated for each of the distorted video sequence as the arithmetic mean of all grades for each sequence (there were 15–22 grades per each video sequence after outlier removal). Every observer graded all degradation types in one session, so scores could not be biased due to the specific degradation tested. Consequently, there was no need for further realignment of the DMOS scores.

To evaluate dependencies between the quality, depth and comfort DMOS grades, Pearson and Spearman's correlation coefficients of these DMOS results were computed. Pearson's correlations between DMOS quality, depth and comfort scores are 0.6418 for quality-depth, 0.6648 for quality-comfort and 0.4957 for depth-comfort. Spearman's correlations between DMOS quality, depth and comfort scores are 0.6417 for quality-depth, 0.7419 for quality-comfort and 0.5927 for depth-comfort. These results show that quality and comfort grades are more similar than other pairs of DMOS scores (quality-depth and depth-comfort). This could be expected as lower quality should in general negatively influence the comfort score, to some level.

6 Evaluation of the results and discussion

6.1 Overall correlation with different objective measures

The next set of results represents Spearman's correlation between 8 full-reference quality measures: PSNR, IQM2 [11], IW-SSIM [40], VIF [36], FI-PSNR [28], MJ3DQA [4], StSD_{lc} [8] (StSD low complexity) and VQM [33] with DMOS results for quality, depth and comfort.

All measures except MJ3DQA were calculated for frame sizes equal to 1920×1080 per view. To speed up calculations, MJ3DQA was calculated for frame size equal to 640×360 obtained from 1920×1080 resolution videos by downscaling by removing 2 pixels out of 3, in each direction, using nearest neighbor interpolation. For the cases of the PSNR, IQM2, IWSSIM and VIF measures, since they are image quality measures and we are evaluating video with two views, the final score was calculated as the average grade between left and right view and then averaged over all 400 frames of the sequence. Concerning the MJ3DQA and FI-PSNR measures, which are 3D image quality measures, the final score was calculated as the average grade across all frames. StSD_lC is a 3D video quality measure and it was calculated using the implementation downloaded from [22], with default values for the parameters. VQM, a 2D video quality measure, was calculated as the average of the left and right view's scores. All the measures were computed on uncompressed sequences (with only degradations 1–17 and 20–22 from Section 3), i.e. the H.264/AVC or vp8 compression used in the grade collection system did not influence these measures results. However, since the PSNR results presented in Section 4 show that only minor change in objective grades can be expected due to the use of high quality compression, this procedure does not introduce perceptible error.

It should also be noted that the measures tested were not designed to evaluate the effects of all the degradations considered in the study, but only for some of them.

For some degradation types, e.g. the frame freezing degradation types 18 and 19, the procedure of averaging over several frames described before is not possible or does not make sense and so these degradations were not subject to further analysis. Overall, objective measures and DMOS scores were compared for 20 degradation types for a total of $20 \times 8 = 160$ DMOS scores.

The values of Pearson's correlation between quality, depth, comfort DMOS scores and PSNR, IQM2, IW-SSIM, FI-PSNR, MJ3DQA, StSD_lC and VQM measures are presented in Table 2. The highest correlations are highlighted in bold. Pearson's correlation was calculated after application of a nonlinear regression step using a 5-parameter logistic function as recommended in [35]:

$$Q(z) = b_1 \cdot \left(\frac{1}{2} - \frac{1}{1 + e^{b_2 \cdot (z - b_3)}} \right) + b_4 \cdot z + b_5 \quad (4)$$

A statistical F-test was performed on each of the two sets of calculated quality measure residuals and the P-value was calculated with 2.5 % significance level, one-tailed test (5 % two-tailed test). In this test the null hypothesis is that data from the two sets follow normal

Table 2 Pearson's correlation between quality, depth, comfort DMOS scores and objective measures

	DMOS_quality	DMOS_depth	DMOS_comfort
PSNR	0.453	0.4238	0.2858
IQM2	0.6095	0.7264	0.5111
IWSSIM	0.5588	0.6887	0.4894
VIF	0.7096	0.6534	0.4113
FI-PSNR	0.5358	0.4064	0.4278
MJ3DQA	0.4539	0.5311	0.4657
StSD_lC	0.3511	0.4945	0.4336
VQM	0.6159	0.691	0.5207

distributions with the same variance, against the alternative that they come from normal distributions with different variances. Results are shown in Table 3. A ‘0’ indicates that means are equal, ‘-’ indicates that the measure in the row has statistically significant lower mean (e.g. better DMOS) than the measure in the column, while ‘+’ means that measure in the row has statistically significant higher mean (e.g. worse DMOS) than the measure in the column. Because the F-test assumes that samples have normal distribution, we also applied a chi-square goodness of fit test against normal distribution, with 5 % significance level, to the tested cases (for quality, depth and comfort). Results are shown in the last column of the Table 3. ‘-’ means that samples do not have normal distribution, while ‘+’ means that samples have normal distribution. It can be seen that some samples do not have normal distribution, so the F-test could give unreliable results in those cases.

The values of Spearman’s correlation between quality, depth, comfort DMOS scores and PSNR, IQM2, IW-SSIM, VIF, FI-PSNR, MJ3DQA, StSD_lc and VQM measures are presented in Table 4. The highest correlations are highlighted in bold.

In some applications objective quality measures are used to estimate the video quality in real time or in near real time and so, besides its quality prediction accuracy, an important factor to consider when choosing a measure is its computational complexity. The computer used in the work was equipped an Intel i7-4770 processor, 16 GB of RAM and running Windows 7. To understand how complex each of the tested measure is, we collected data on their computation complexity, presented in Table 5 which shows computation time per each frame, both left and right.

It can also be observed that all tested measures exhibited lower correlation with quality scores, than with depth or comfort DMOS scores. It is worth noticing that all Spearman’s correlations are lower than 0.5 evidencing weak relationships between the grades under analysis.

For some degradations very low correlation values are expected because the objective quality measures studied were not designed to accurately model the effect of those types of impairments. Such are the cases for instance of errors in horizontal or vertical disparity, frame rate reduction and 3D-to-2D-to-3D switching. According to the authors’ knowledge, none of the objective measures (image, video, 3D image or especially 3D video) described in the scientific literature have been designed for all of the mentioned degradation types. Such degradation types are specific in the sense that they can affect differently the quality, depth and comfort of the 3D video (e.g. resulting in good overall quality but very low comfort).

The low values of Pearson’s and especially Spearman’s correlation are a clear indicator that new 3D objective quality measures should be researched and proposed, as the measures

Table 3 Pearson’s correlation and statistical significance

	PSNR	IQM2	IWSSIM	VIF	FI-PSNR	MJ3DQA	StSD_lc	VQM
PSNR	0,0,0	0,+0	0,+0	+,+0	0,0,0	0,0,0	0,0,0	0,+0
IQM2	0,-0	0,0,0	0,0,0	0,0,0	0,-0	0,-0	-,-0	0,0,0
IWSSIM	0,-0	0,0,0	0,0,0	+0,0	0,-0	0,0,0	0,-0	0,0,0
VIF	-,-0	0,0,0	-0,0	0,0,0	-,-0	-0,0	-0,0	0,0,0
FI-PSNR	0,0,0	0,+0	0,+0	+,+0	0,0,0	0,0,0	0,0,0	0,+0
MJ3DQA	0,0,0	0,+0	0,0,0	+0,0	0,0,0	0,0,0	0,0,0	0,+0
StSD_lc	0,0,0	+,+0	0,+0	+0,0	0,0,0	0,0,0	0,0,0	+,+0
VQM	0,-0	0,0,0	0,0,0	0,0,0	0,-0	0,-0	-,-0	0,0,0
chi2	+,+,+	+,+,-	+,+,-	-,+,+	+,+,+	+,+,+	+,+,-	+,+,+

Table 4 Spearman's correlation between quality, depth, comfort DMOS scores and objective measures

	DMOS_Quality	DMOS_Depth	DMOS_Comfort
PSNR	0.2036	-0.0774	-0.228
IQM2	-0.2747	-0.3925	-0.4726
IWSSIM	-0.0661	-0.3442	-0.4268
VIF	-0.3382	-0.232	-0.3569
FI-PSNR	0.2427	0.013	-0.1795
MJ3DQA	-0.0277	-0.16	-0.4116
StSD_lc	0.2557	0.3902	0.4611
VQM	0.2921	0.3782	0.4977

obtained by extending 2D measures to evaluate 3D video do not perform well. It was also observed that the 3D measures analysed (image and video) did not perform well either. New and better measures have to be designed with particular care on what concerns their computation complexity and performance on stereoscopic video affected by different types of degradation, both induced by compression and processing operations (symmetric and asymmetric H.264/AVC, H.265/HEVC compression of 3D content) as well as transmission impairments (like packet losses, streaming freezes, 3D to 2D switches, etc.) as well as content impairments (horizontal and vertical disparity, frame rate reduction, geometric distortions etc.).

In the next subsections we will examine correlation between objective measures and some specific subsets of 3DVCL@FER database. Subsection 6.2 will examine compression related degradations and resizing (subset 1) to be able to compare our results with another 3D video database. Subsection 6.3 will examine compression related degradations, resizing and packet losses (subset 2), subsection 6.4 temporal degradations (subset 3), subsection 6.5 horizontal and vertical disparity (subset 4) and subsection 6.6 will propose some other possible subsets of 3DVCL@FER database.

6.2 Comparison with other 3D databases: compression related degradations and resizing degradation (subset 1)

To compare our results with some of the existing 3D databases, we used data from the publicly available NAMA3DS1-COSPAD1 database [37] (later called Nantes database) and compared the correlation with objective measures (mentioned earlier) between this database and a subset

Table 5 Timing (milliseconds per each frame, both left and right) for all tested objective measures

	Time (ms)
PSNR	327.4
IQM2	1384.1
IWSSIM	4825.5
VIF	12499
FI-PSNR	1632.9
MJ3DQA	22738
StSD_lc	879.7
VQM	3258

of the 3DVCL@FER database. This subset (subset 1) consists of resize degradations (number 2 in section 3), H.264/AVC degradation (number 11 and 12 in section 3) and jpeg2000 degradation (number 17 in section 3) and includes 32 DMOS results. The NAMA3DS1-COSPAD1 database is made up of 10 different reference sequences (8 of which are the same as in 3DVCL@FER) and 100 DMOS results for degraded sequences: 30 for H.264/AVC compression, 40 for jpeg2000 compression and 30 non-compression related artifacts (resizing, edge enhancement and combination). Pearson's correlation between 32 DMOS quality scores from 3DVCL@FER and Nantes database is 0.9624, showing that similar DMOS were obtained in the evaluations carried in both laboratories. Results for objective measures are presented in Table 6. The highest correlations are highlighted in bold.

From Table 6 it can be concluded that we obtained correlations between subjective scores and objective measures similar to those obtained in the Nantes experiments. Minor differences may be due to the different number of tested video sequences (32 from 3DVCL@FER versus 100 from Nantes). The best performing measure among the tested measures is IWSSIM for DMOS depth and DMOS comfort scores in 3DVCL@FER database and in Nantes database. VIF measure is best performing measure for DMOS quality scores in 3DVCL@FER database, however IWSSIM has nearly similar correlation (and is the second best in this case). VIF measure is also second best measure in Nantes database. Also, this means that the ACR-HR subjective comparison method is reliable for use with these types of degradations, when testing 3D video sequences, as it gave led to similar conclusions in different laboratories for both subjective and objective assessment. The quality, depth and comfort DMOS scores of subset 1 show high Pearson's correlation with a value of 0.9275 for quality-depth, 0.9503 for quality-comfort and 0.9088 for depth-comfort. Also in this case the highest correlation is observed between quality and comfort DMOS scores (a similar conclusion was drawn in Section 5).

6.3 Comparison between objective measures using compression related, resizing and packet losses degradations (subset 2)

In this section we will present correlation between earlier mentioned objective quality measures and degradations (subset 2): 11, 12, 13, 14, 17, 21 and 22 (compression related), 2 (resizing) and 15 (packet losses) described in Section 3. Overall this subset consists of 72 DMOS grades. This subset could represent typical degradation types from transmitter to the

Table 6 Spearman's correlation in subset 1 of 3DVCL@FER and Nantes database, between objective measures and DMOS

	3DVCL@FER			NANTES
	DMOS_quality	DMOS_depth	DMOS_comfort	DMOS
PSNR	-0.7346	-0.7636	-0.7518	0.7258
IQM2	-0.8581	-0.8893	-0.8464	0.8781
IWSSIM	-0.8878	-0.9208	-0.8922	0.9338
VIF	-0.9021	-0.8735	-0.8611	0.9280
FL-PSNR	-0.6793	-0.6866	-0.687	0.7191
MJ3DQA	-0.7614	-0.7254	-0.7452	0.7249
StSD_lc	0.809	0.7896	0.827	-0.8203
VQM	0.8501	0.8361	0.8618	-0.8601

receiver: compression related on the receiver side, packet losses in the transmission channel and resizing on the receiver side (e.g. upsampling to full HD resolution). Results are presented in Table 7. The highest correlations are highlighted in bold.

Overall correlation is generally lower than in Section 6.2 because of the more different types of degradations tested. From the results it can be concluded that for subset 2 the best performing measure among those tested is IWSSIM for all DMOS scores, while IQM2 is second best for DMOS_quality and DMOS_comfort scores (StSD_lc is second best for DMOS_depth scores). StSD_lc also produces comparable results with IWSSIM and IQM2 measures. StSD_lc has also lower calculation time than IWSSIM or IQM2, as shown in Table 5, so it is the most useful objective metric for this subset of degradations. Pearson's correlation in subset 2 between DMOS quality, depth and comfort scores is: 0.8905 (quality-depth), 0.9554 (quality-comfort) and 0.8663 (depth-comfort).

6.4 Temporal degradations (subset 3)

We have also studied subjective grade variations for temporally induced degradation types (subset 3): 3 (frame rate reduction), 18 (frame freezing, online streaming), 19 (frame freezing, offline streaming) and 20 (3D-2D-3D switching). Those degradations could represent typical visual impairments occurring due to transmission channel problems coupled with decoder-side missing frame interpolation. The correlation between the grades in subset 3 and tested objective measures are not presented because it is not clear how to model the quality drop due to frame freezes.

Average DMOS for all 8 sequences per degradation type is shown in Table 8. Statistical significance for DMOS (quality, depth, comfort), presented in Table 9, was computed using two sample t-test, with 2.5 % significance level, one-tailed test. This test has as null hypothesis that data in the tested samples (DMOS scores in our case) are independent random samples from normal distributions with equal means and equal but unknown variances, against the alternative that the means are not equal. Results are shown in '0' means that means are equal, '-' means that case in related row has statistically significant lower mean (e.g. better DMOS) than case in related column, while '+' means that case in related row has statistically significant higher mean (e.g. worse DMOS) than case in related column.

From Table 9 it can be concluded that 3D-2D-3D switching (degradation type 20) has the best and statistically more significant average DMOS quality. Second best is frame rate reduction degradation (type 3). Online and offline frame freezing degradations (types 18 and 19) have statistically similar average DMOS quality scores, the lowest in subset 3. Average

Table 7 Spearman's correlation in subset 2 of 3DVCL@FER database between objective measures and DMOS

	DMOS_quality	DMOS_depth	DMOS_comfort
PSNR	-0.682	-0.7133	-0.719
IQM2	-0.7898	-0.8246	-0.769
IWSSIM	-0.8041	-0.8627	-0.7927
VIF	-0.7206	-0.7944	-0.6386
FL-PSNR	-0.7	-0.626	-0.7263
MJ3DQA	-0.6434	-0.6093	-0.646
StSD_lc	0.7836	0.84	0.7686
VQM	0.7824	0.7984	0.7746

Table 8 Average DMOS scores for tested subset 3 of 3DVCL@FER database

Degradation type	DMOS_quality	DMOS_depth	DMOS_comfort
3	62.2156	49.2634	50.4179
18	67.2874	49.784	53.0292
19	67.1778	47.8073	52.3626
20	57.0644	51.4516	51.5941

DMOS depth and DMOS comfort scores were also compared and are also shown in Table 9. DMOS average depth scores are statistically significant only between degradation 19 (frame freezing, offline streaming) and 20 (3D-2D-3D switching), where in this case DMOS depth scores are lower, as expected. DMOS average comfort scores are statistically insignificant in this subset.

Related conclusions were presented in [2], where observers preferred 3D uncompressed sequences with higher frame rate (48 or 60 fps) over 3D uncompressed sequences with lower frame rate (24 or 30 fps), especially for sequences with higher motion levels. In our experiment 25 fps original sequences were used, which were degraded to 1/3 of the original fps. In Fig. 6 sequences (1–8) with higher motion levels (Temporal information parameter) generally show worse DMOS quality and worse comfort scores, than sequences with lower motion levels. Spearman's correlation between temporal information (left view) and DMOS quality score (for degradation type 3) is 0.4286, while Spearman's correlation between temporal information (left view) and DMOS comfort score (for degradation type 3) is 0.6905. Spearman's correlation between temporal information (left view) and DMOS depth score (for degradation type 3) is 0.1905, which could be expected (this means that frame rate reduction has very low influence on DMOS depth scores).

6.5 Horizontal and vertical disparity (subset 4)

In subset 4, we tested subjective grade variations for horizontal and vertical disparity: degradation types 6–7 (horizontal disparity) and 8 (vertical disparity). Average DMOS for all 8 sequences per degradation type is shown in Table 10. Statistical significance for DMOS (quality, depth, comfort), presented in Table 11, was computed using two sample t-test, with 2.5 % significance level, one-tailed test, similarly to the subsection 6.4. Results clearly show that vertical disparity has much higher impact on final quality, depth and especially comfort scores, e.g. all DMOS scores are worse for vertical disparity. Also, it can be concluded that horizontal disparity has the same impact for degradation types 6 (30 pixels shift inwards) and 7 (30 pixels shift outwards).

Table 9 Two sample t-test between DMOS (quality, depth, comfort) scores for tested subset 3 of 3DVCL@FER database

Degradation type	3	18	19	20
3	0,0,0	-,0,0	-,0,0	+,0,0
18	+,0,0	0,0,0	0,0,0	+,0,0
19	+,0,0	0,0,0	0,0,0	+,-,0
20	-,0,0	-,0,0	-,+,0	0,0,0

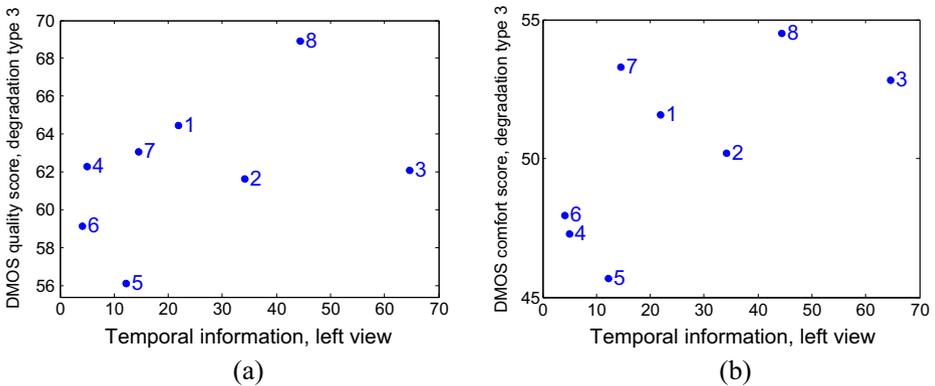


Fig. 6 Temporal information, *left* view, compared with DMOS scores for frame rate reduction degradation (type 3): (a) DMOS quality, (b) DMOS comfort

6.6 Other possible uses of the 3DVCL@FER database

From the exposition in section 6.1 it can be concluded that all tested objective measures (2D image, 3D image and 3D video) have failed to predict accurately DMOS grades. Probably this poor performance is due to a mismatch between the applicability conditions of the measures and the type of degradations specific to 3D video (horizontal and vertical disparity, 3D-to-2D-to-3D switching, frame rate reduction and frame freezing of 3D content), which were used to assemble the database described in this paper. This means that more effort should be given towards development of objective measures for 3D specific degradations.

From the analyses presented in sections 6.2 and 6.3, it can also be concluded that higher correlation was obtained using 2D image quality measures, than when using 3D image or 3D video quality measures. Most likely this occurred because most of the degradations represented in the database described were symmetrically induced (except degradation 13).

New 3D objective video quality measures could be developed based on the 3DVCL@FER database, either from one part or the whole database. Obviously, such measures should take into account all 3 values that are connected with 3D video evaluation: quality, depth and comfort. In some cases, those values can be similar (subsets 1–2), but overall they represent 3 different data about the final subjective evaluation (subsets 4, 5 and all database). One possible approach could be a combination of existing measures, each for one score (e.g. VIF for quality, IQM2 for depth and VQM for comfort score), as they gave best correlation in the whole database. Measures should be also probably adjusted to take into account degradations specific to 3D video.

Table 10 Average DMOS scores for tested subset 4 of 3DVCL@FER database

Degradation type	DMOS_quality	DMOS_depth	DMOS_comfort
6	49.6095	45.6089	45.8157
7	48.4194	44.1713	48.0076
8	55.2073	50.7748	63.4968

Table 11 Two sample t-test between DMOS (quality, depth, comfort) scores for tested subset 4 of 3DVCL@FER database

Degradation type	6	7	8
6	0,0,0	0,0,0	-, -, -
7	0,0,0	0,0,0	-, -, -
8	+, +, +	+, +, +	0,0,0

Different subsets of sequences and respective grades can be extracted from the 3DVCL@FER database that are matched to the mix of degradations expected to be observed in a hypothetical 3D video coding and transmission system. Those subsets can be then used as training or testing datasets for newly created objective measures tailored to the same mix of impairments. The database content includes enough types of degradations to allow building very specific subsets. As an example, using the data for degradations 4 through 9 (inter-camera luminance and disparity errors), it is possible to create a subset of sequence and subjective grades to evaluate the performance of objective measures designed to measure the effect of incorrect camera settings.

As a possible future research activity one could also perform an evaluation of the same dataset using web-based systems such as [6] or [7] and compare the grades collected with the grades we obtained in both the Portuguese and Croatian laboratories.

7 Conclusion

In this paper we have presented the 3DVCL@FER 3D video sequence database. It consists of 8 original 3D video sequences and 22 degraded 3D video sequences for each original sequence, with various degradation types including degradations specific to 2D or 3D stereoscopic systems.

The results from subjective assessments of 3D video quality on the 3DVCL@FER database conducted in two laboratories in Croatia and Portugal were compared with several objective methods using figures of merit like the Spearman correlation coefficient.

Our findings show that the correlation between subjective grades and objective quality estimation methods for 3D video is still inadequate, especially when comparing widely different degradation types. New objective methods are being developed which hopefully are better adapted to 3D video quality assessment. The research presented in this paper can be used to test and evaluate new 3D video quality estimation methods as established 3D video sequence databases contain different types of distortion and subjective assessment grades.

Future research could be also directed towards comparison of MOS and DMOS quality scores collected using pure web-based setups and scores obtained in more controlled laboratory based grading sessions.

References

1. Aflaki P, Hannuksela MM, Hakala J, Häkkinen J, Gabbouj M (2011) Estimation of subjective quality for mixed-resolution stereoscopic video. 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON) 1–4

2. Banitalebi-Dehkordi A, Pourazad MT, Nasiopoulos P (2015) The effect of frame rate on 3D video quality and bitrate. *3D Res* 6(1):1–13. doi:10.1007/s13319-014-0034-3
3. Battisti F, Bosc E, Carli M, Le Callet P, Perugia S (2015) Objective image quality assessment of 3D synthesized views. *Signal Process Image Commun* 30:78–88. doi:10.1016/j.image.2014.10.005
4. Chen M-J, Su C-C, Kwon D-K, Cormack LK, Bovik AC (2013) Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Process Image Commun* 28(10):1143–1155. doi:10.1016/j.image.2013.05.006
5. Cheng E, Burton P, Burton J, Joseski A, Burnett I (2012) RMIT3DV: pre-announcement of a creative commons uncompressed HD 3D video database. *Qual Multimed Exp (QoMEX 2012)* 212–217
6. crowd3d.co.it.pt/suis3d. Accessed 21 July 2015
7. crowd3d.co.it.pt/suis3d_webm. Accessed 21 July 2015
8. De Silva V, Arachchi HK, Ekmekcioglu E, Kondoz A (2013) Towards an impairment metric for stereoscopic video: a full-reference video quality metric to assess compressed stereoscopic video. *IEEE Trans Image Process* 22(9):3392–3404. doi:10.1109/TIP.2013.2268422
9. Dumic E, Grgic S, Jiménez Bernejo D, Silva Cruz LA (2014) Benchmark on state of the art objective measures for 3D stereoscopic video quality assesment on the Nantes database. *Proc. ELMAR-2014* 119–123
10. Dumic E, Grgic S, Grgic M (2010) Comparison of HDTV formats using objective video quality measures. *Multimed Tools Appl* 49(3):409–424. doi:10.1007/s11042-009-0441-2
11. Dumic E, Grgic S, Grgic M (2014) IQM2 - New image quality measure based on steerable pyramid wavelet transform and structural similarity index. *SIViP* 8(6):1159–1168. doi:10.1007/s11760-014-0654-3
12. ftp://ftp.ivec.polytech.univ-nantes.fr/NAMA3DS1_COSPAD1/Avi_videos/HRC_00_Reference/. Accessed 21 July 2015
13. Goldmann L, De Simone F, Ebrahimi T (2010) A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. *Electronic Imaging (EI), 3D Image Processing (3DIP) and Applications, SPIE Proceedings* 7526
14. <http://avisynth.nl/index.php/Toon>. Accessed 21 July 2015
15. <http://ffmpeg.zeranoe.com/builds/>. Accessed 21 July 2015
16. <http://goo.gl/wSEkHd>. Accessed 21 July 2015
17. <http://hevc.kw.bbc.co.uk/git/wjctvc-3de.git/commit/d85c6b6e015c86c5c7c99ca9983304c14f8d9ad1>. Accessed 21 July 2015
18. <http://iphone.hhi.de/suehring/tml/download/>. Accessed 21 July 2015
19. <http://sourceforge.net/projects/avisynth2/>. Accessed 21 July 2015
20. <http://www.cl.cam.ac.uk/~mgk25/download/stirmark-1.0.tar.gz>. Accessed 21 July 2015
21. http://www.kakadusoftware.com/index.php?option=com_content&task=view&id=26&Itemid=22. Accessed 21 July 2015
22. http://www.mathworks.com/matlabcentral/fileexchange/42106-stereoscopic-structural-distortion-metric-stds/-content/StSD_lc.m. Accessed 21 July 2015
23. <http://xstreamer.atlantis.ugent.be/>. Accessed 21 July 2015
24. ITU-R BT.2021 (2012) Subjective methods for the assessment of stereoscopic 3DTV systems. International Telecommunication Union/ITU radiocommunication sector
25. ITU-R BT.500-13 (2012) Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union/ITU radiocommunication sector
26. ITU-T Recommendation P.910 (2008) Subjective video quality assessment methods for multimedia applications
27. Joveluro P, Malekmohamadi H, Fernando WAC, Kondoz AM (2010) Perceptual video quality metric for 3D video quality assessment. *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)* 1–4
28. Lin Y, Wu J (2014) Quality assessment of stereoscopic 3D image compression by binocular integration behaviors. *IEEE Trans Image Process* 23(4):1527–1542. doi:10.1109/TIP.2010.2092435
29. Lopez JP, Rodrigo JA, Jimenez D, Menendez JM (2013) Stereoscopic 3D video quality assessment based on depth maps and video motion. *EURASIP J Image Video Process* 62:1–14. doi:10.1186/1687-5281-2013-62
30. Ma R, Cheung N-M, Au OC, Tian D (2013) Novel distortion metric for depth coding of 3D video. *2013 International Conference on Image Processing (ICIP)* 1714–1718

31. Moorthy AK, Su C-C, Mittal A, Bovik AC (2013) Subjective evaluation of stereoscopic image quality. *Signal Process Image Commun* 28(9):870–883. doi:10.1016/j.image.2012.08.004
32. Mysirlidis C et al. (2015) STESCAL3D: subjective evaluation of HD stereo video streaming using H.264 SVC in diverse laboratory environments. *Qual Multim Exp (QoMEX 2015)* 1–6
33. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. *IEEE Trans Broadcast* 50(3):312–322
34. QUALINET Multimedia Databases v5.5, http://dbq-wiki.multimediatech.cz/_media/qi0542.pdf. Accessed 21 July 2015
35. Sheikh HR (2004) Image quality assessment using natural scene statistics. Ph.D. dissertation, University of Texas at Austin
36. Sheikh HR, Bovik AC (2006) Image information and visual quality. *IEEE Trans Image Process* 15(2):430–444. doi:10.1109/TIP.2005.859378
37. Urvoy M, Barkowsky M, Cousseau R, Koudota Y, Ricorde V, Le Callet P, Gutierrez J, Garcia N (2012) NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. *Quality of Multimedia Experience (QoMEX 2012)* 109–114
38. VQEG 3DTV Group (2012) Test plan for establishing a ground truth for quality of experience in 3D for assessment methodologies in 3D video quality assessment, GroTruQoE3D1, Draft Version 1.0
39. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Proc* 13(4):600–612. doi:10.1109/TIP.2003.819861
40. Wang Z, Li Q (2011) Information content weighting for perceptual image quality assessment. *IEEE Trans Image Process* 20(5):1185–1198. doi:10.1109/TIP.2010.2092435
41. Xing L, You J, Ebrahimi T, Perks A (2013) Stereoscopic quality datasets under various test conditions. *Qual Multim Exp (QoMEX 2013)* 136–141
42. Zota V, Janssen J-K (2010) Tiefenbehandlung, 2D-Videos in 3D abspielen, c't Magazin für Computer Technik, Heise Verlag 6:116



Emil Dumic is an Assistant Professor at the University North, Department of Electrical Engineering. He received his PhD degree from University of Zagreb, Faculty of Electrical Engineering and Computing in 2011. He is the author or coauthor of 24 conference papers, 8 journal papers and has written one book chapter. His current research interests include development of objective image and video quality measures, subjective and objective assessments on image, video and 3D video databases, image interpolation, impact of different channel models on bit error rate in different DVB standards, etc.



Sonja Grgic received the B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 1989, 1992 and 1996, respectively. She is currently full Professor at the Department of Wireless Communications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. Her research interests include television signal transmission and distribution, picture quality assessment and wavelet image compression. She has participated in 10 domestic and international scientific projects. She was a project leader of the research project “Picture Quality Management in Digital Video Broadcasting” financed by the Ministry of Science, Education and Sports of the Republic of Croatia. She is author or co-author of 20 papers published in scientific journals, more than 120 papers published in conference proceedings of international scientific conferences as well as of 15 reviewed studies and expert works. She was editor of 6 international conference proceedings. She is a member of IEEE and SMPTE.



Kresimir Sakic was born in 1983 in Zagreb. He attended elementary school in Zagreb, where he also finished high school in 2001. The same year he started his studies at the Faculty of electrical engineering and computing at University of Zagreb. He graduated in December 2006, defending the graduate thesis “Transmission error protection procedures in the DVB-H system” under the supervision of Prof. Mislav Grgic. In March 2007 he started working as a Broadcasting Planning Expert in the Croatian Regulatory Authority for Network Industries, dealing mostly with frequency planning, interference analysis, and licensing/tendering procedures of television broadcasting networks, as well as with cross-border coordination. He was also engaged in work of the CEPT working groups TG4 (Digital Dividend) and TG6 (Long Term Vision UHF) and other activities related to the allocation of the first and second Digital Dividend. In the year 2008 he began postgraduate study in the University of Zagreb, Faculty of Electrical Engineering and Computing under the supervision of Prof. Sonja Grgic. He has published numerous scientific conference papers in conference proceedings, and participated in the organisation and review processes of several scientific conferences (ELMAR, IWSSIP). He also deals with a bit of sports, especially mountain biking. He speaks English and German fluently.



Pedro Miguel Regalo Rocha was born in Coimbra, Portugal, in 1991. He is currently finishing the Master's degree in Electrical and Computer Engineer at University of Coimbra. Also, he was a researcher at Instituto de Telecomunicações in Coimbra, Portugal.



Luis A. da Silva Cruz received the Licenciado and M.Sc. degrees in Electrical Engineering from the University of Coimbra, Portugal, in 1989 and 1993 respectively. He also holds an MSc degree in Mathematics and a Ph.D. degree in Electrical Computer and Systems Engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, US granted in 1997 and 2000 respectively. He has been with the Department of Electrical and Computer Engineering of the University of Coimbra in Portugal since 1990 first as a Teaching Assistant and as an Assistant Professor since 2000. He is a researcher of the Institute for Telecommunications of Coimbra where he works on video processing and coding, medical image processing and wireless communications. He is a member of the SPIE and IEEE technical societies.