Crowdsourced Subjective 3D Video Quality Assessment

[1]Emil Dumic, [2]Kresimir Sakic, and [3]Luis A. da Silva Cruz

[1]University North, Department of Electrical Engineering, 104. brigade 3, 42000 Varaždin, Croatia; ORCID: orcid.org/0000-0002-0262-5595
[2]Croatian Regulatory Authority for Network Industries (HAKOM), Radio Communication Department, Roberta Frangeša Mihanovića 9, 10110 Zagreb, Croatia
[3]Instituto de Telecomunicações and Department of Electrical and Computer Engineering, University of Coimbra, Rua Sílvio Lima, Pólo II, 3030-290, Coimbra, Portugal; ORCID: orcid.org/0000-0003-1141-4404

email: emil.dumic@unin.hr

***Abstract* - This article proposes a new method for subjective 3D video quality assessment based on crowdsourced workers – *Crowd3D*. The limitations of traditional laboratory-based grade collection procedures are outlined, and their solution through the use of a crowd-based approach is described. Several conceptual and technical requirements of crowd-based 3D video quality assessment methods are identified and the solutions adopted described in detail. The system built takes the form of a web-based platform that supports 3D video monitors, and orchestrates the entire process of observer validation, content presentation and quality, depth and comfort grade recording in a remote database. The crowdsourced subjective 3D quality assessment system uses as source contents a set of 3D video and grades database assembled earlier in a laboratory setting. To evaluate the validity of the crowd-based approach the grades gathered using the crowdsourced system were analysed and compared to a set of grades obtained in laboratory settings using the same dataset. Results show that it is possible to obtain Pearson's and Spearman's correlation up to 0.95 for quality DMOS (Difference Mean Opinion Score) and 0.96 for quality MOS (Mean Opinion Score), when comparing with laboratory grades. Apart from the present study, the 3D video quality assessment platform proposed can be used with advantage for further related research activities, reducing the time and cost compared to traditional laboratory-based quality assessments.**

***Index Terms*—crowdsourcing; 3D video quality; subjective assessment;Crowd3D**

I. INTRODUCTION

Research on 2D and 3D video processing, coding and quality modelling often requires access to video clips annotated with grades representing human opinion of their quality. Usually these grade datasets are compiled by enrolling subjects to participate in video quality assessments campaigns, during which they watch a number of video sequences and rate their quality on either an absolute or a relative scale following one of the protocols defined in specialized recommendations. One such recommendation is ITU-R BT.500-13 [1], with a scope of application limited to 2D video contents. An extension of this protocol to stereoscopic 3DTV systems has been developed by ITU and is available as recommendation ITU-R BT.2021 [2], to which three other 3D video related recommendations have been added recently; ITU-R P.914 (Display requirements for 3D video quality assessment) [3], ITU-R P.915 (Subjective assessment methods for 3D video quality) [4] and ITU-R P.916 (Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video) [5]. In subjective 2D video quality assessment the observers rate video on a single dimension that quantifies quality [6], [7], but in subjective 3D video quality assessment other quality indicators specific to 3D such as depth quality and visual comfort have to be rated as well. That means that for each 3D video sequence, test subjects have to indicate three different grades, as opposed to one in the 2D video case, thus making the evaluation procedure longer and more cumbersome, and more prone to inter and intra-observer variability. To improve the reliability of the quality grades collected in subjective 3D video quality assessments, the test subjects need to pass a set of stereoscopic vision screenings, alongside colour and vision acuity tests, thus accruing to the logistic complexities of 3D video evaluations and their costs. In the past these constraints have put practical limits on 3D video quality subjective quality grade collection, both to the amount of grade data collected as wells as to diversity of the grade sources, most times limited to academic and industrial research-laboratory settings.

Recent developments on crowdsourced image [8], [9] and video quality assessment [10] and the availability of crowdsourcing platforms such as Microworkers [11] and Amazon Mechanical Turk [12] have provided an alternative to laboratory-based quality evaluations. Using crowdsourced evaluators it is now possible to have 3D video quality assessments done by many observers at multiple locations, extending the evaluators recruitment domain and thus solving one of the problems of this type of studies, the assembly of a diversified medium to large set of graders. However the geographical distribution of observers together with the diversity of their backgrounds and other specificities of this type of grade collection modus introduce several new technical and conceptual challenges that need to be solved before crowdsourced 3D video quality assessment campaigns are effective and their results trustworthy.

In the following sections these challenges will be identified and corresponding solutions will be described, resulting in a set of procedures and tools which form the proposed framework of the new method for crowdsourced subjective 3D video quality assessment – *Crowd3D*. The system described was developed as a web-based platform that controls several stages of the

evaluation sessions performed remotely by (crowd) workers, including several types of verifications, and finally the grade collection.

The system proposed was used to gather subjective quality ratings for 3D video sequences that were prepared for and used in a previous study done in a laboratory setting, that resulted in a grade annotated 3D video database, 3DVCL@FER, as reported in [13]. The quality, depth and comfort grades collected using the proposed crowd-based platform were compared to the laboratory-based grades. The grades obtained were subject to extensive analysis which enabled drawing conclusions about the feasibility and reliability of the procedure proposed. It will be shown that correlation between overall quality scores with laboratory evaluation will be high, while depth and comfort scores will be somewhat lower. One of the possible problems may be lower number of overall evaluations per video sequence – around 34.8 (in crowdsourced environment usually it is possible to quickly collect several hundred grades or samples), because nowadays people still do not usually have 3D equipment. Still, this was enough for quality grades, but not for comfort or depth grades. Another problem may be different environmental settings that have influence on comfort ratings, while e.g. different distances from monitors may have different perceived depth, and those factors cannot be strictly controlled in crowdsourced evaluation.

The remainder of the text is organized as follows. Section II summarizes previous works on related themes. The details about the web-based application and test setup used for this crowdsourced 3D video quality assessment project are exposed in Section III. Section IV reports on the individual quality grades collected and a grade comparison study between crowdsourced and laboratory-based subjective 3D video quality assessments. Section V discusses about results obtained from crowdsourced experiment, as well as about comparisons with results from other subjective 3D video quality assessments. Finally, Section VI presents our conclusions.

II. RELATED WORK

To the authors best knowledge, there are no published results on crowdsourced 3D video quality assessment where the assessments were conducted using 3D displays (3D monitors or 3D TV sets), although there is some work concerned with crowd-based quality assessment of multiview video plus depth coding like [14], where Hanhart et. al. investigated two possible approaches to crowd-based quality assessment of multiview video plus depth (MVD) content presented on 2D displays. Another work used subjective 3D video quality assessments to build the 3DVCL@FER [13] 3D video database annotated with quality grades. Despite its attractiveness as a way to quickly gather large numbers of quality grades at low cost, crowdsourced 3D video quality assessment faces a number of technical and conceptual challenges as described later.

Hoßfeld et. al. have shown that a two-stage design can assemble a pseudo-reliable user pool with specific characteristics or user equipment [15], [16]. Stage one should be very short and would serve only to select users that have normal stereoscopic vision (i.e. are able to perceive depth) and a 3D monitor or TV set. Only the participants who pass stage one should be allowed to take part in stage two. In stage two the actual crowdsourced subjective 3D video quality assessment is done.

In [17], experiments are described that test "perceived depth", "perceived image quality" and "perceived naturalness" in images with different levels of blur and different depth levels. Conclusion was that naturalness incorporates both blur level as well as depth level, while image quality does not include depth level, thus naturalness is a more promising concept.

In [18] authors proposed 3D Quality Model based on weighted sum of perceived image quality and perceived depth. Adding blur or noise may affect both perceived image quality and perceived depth. In [19] authors describe visual discomfort in stereoscopic displays and different factors that can affect it: excessive binocular disparity, accommodation and convergence mismatch, (un)comfortable viewing distances, stereoscopic distortions.

Comparison between different subjective 3D video quality evaluations has been presented in [14] (between crowd-based and lab-based test; authors compared MOS quality scores, using video sequences coded with MVC+D and 3D-AVC, with different bitrate, and converted to different synthesized views for subjective test), [20] (3 different laboratories; similar setup for tests as in [14]) and [21] (3 different laboratories; authors tested 10 degradation types from NAMAS1-COSPAD dataset [22]). In [23] authors carried an experiment to determine the impact of certain video characteristics such as fast in-scene motion, large changes in disparity and depth discontinuities caused by subtitles, in terms of visual comfort via different measurement methods. An analysis of the continuous assessment scores (tested sequences were two 3-D movies of approximately 15 min each, both with and without subtitles) revealed that visual comfort could be predicted from a linear combination of these video characteristics per scene.

In [24] authors presented a new method to quantify stereoscopic visual performance at different base disparity levels inside and outside the zone of comfortable viewing, which could allow to adjust individual zones of comfortable viewing (e.g. using this approach, users could automatically and individually adjust settings for a 3D television consumption).

In [25] authors presented a novel framework for jointly modeling QoE and user behavior, where user behavior is treated as one of the framework dimensions along with system performance and user state. It can be used for traditional QoE, user behavior, charging and pricing models over churn issues and the impact of user characteristics, problems related to energy consumption etc.

III. Application Design and Test Setup

A. Problem description and challenges

Although crowdsourced tests generally reduce the time and cost compared to traditional laboratory-based quality assessments, crowdsourced 3D video quality assessment faces different technical and conceptual challenges. The main technical challenges are internet access bandwidth and quality constraints, support of user equipment to display the required stimuli and lack of information about the viewing environment where the crowdsourced subjective 3D video quality assessment takes place.

The second important challenge is the support of different types of user equipment to display the required stimuli. This challenge has implications that translate into more demanding hardware and software requirements. On the hardware side, the most important requirement is that the users must have a 3D monitor or 3D TV set capable of displaying the 3D video sequences. On the software side, because the availability of high-end user equipment cannot be readily assumed, optimisation for smooth execution on older computers is needed.

Another important challenge is the trustworthiness of the user and user data. Commercial crowdsourcing platforms such as Microworkers [11] and Amazon Mechanical Turk [12] have a large pool of diverse workers and implement a worker rating scheme based on the success rate of finished jobs. The existence of dishonest users on the commercial crowdsourcing platforms means that additional reliability mechanisms (later called ARMs) need to be implemented. Those ARMs can be implemented before, during and after the crowdsourced subjective 3D video quality assessment test campaign. ARMs can be used during the quality assessment sessions to identify unreliable users and dismiss their results. After this step a crowdsourced subjective 3D video quality assessment test campaign can be conducted including the application of the recommended statistical analysis, as will be described later on. Because of the requirement of access to a 3D monitor or 3D TV set a two-stage crowdsourcing test campaign is preferred.

To make the system usable by a large number of crowdworkers, it should be designed to use standard browsers, not requiring any special plugins. The content to be evaluated should be easy to download, using near-lossless compression and should be pre-stored in the browsers cache to avoid playback interruptions. For the purpose of our research on crowdsourced subjective 3D video quality assessment, a web-based application was developed following the tenets enunciated in the previous section. Although, the application could be run in either Google Chrome [26] or Mozilla Firefox browser [27], for the crowdsourced assessments Google Chrome and x264 encoded video sequences were used, as this test setup does not require any additional software installation from the user side. Because of the complexity of the crowdsourced subjective 3D video quality assessment procedure, dedicated test server in Portugal was used on high-speed network, running *Apache v2.4.23* and *PHP v5.5.38*.

All the additional reliability mechanisms (ARMs) implemented in the web-based application are listed below:

a) Forcing the browser into full screen mode during the whole assessment procedure. If the user tries to exit the full screen mode an error message is displayed, the test procedure is stopped and the start page is loaded;

b) GUI is rendered in 3D mode;

c) To prevent the hasty scoring the users are not allowed to submit a score before a predetermined amount of observation time has elapsed; this guard time was set to 5 seconds;

d) Application level monitoring of the results, web browser type/version, screen resolution and operating system is used. The default rating count and average grades of original 3D video sequences results are monitored. If the users choose more than five default ratings (they do not move the rating sliders for 5 3D video sequences) they are marked as "potential cheater" in the results database. They are marked the same way if the average grades of original 3D video sequences is below 1.5. Default position was set at the middle of the scale. Setting it at either end of the scale or at an invalid position (which would then start at 0 when moved) could bias the scores of the observers towards the ends of the scale;

e) Context and demographic monitoring are implemented through a questionnaire where users are asked to provide information about their 3D monitor type, illumination type, time of day, age, gender and country. Most of the questions are implemented through drop-down menus so that the users do not spend a lot of time filing out the questionnaire. Those questions are answered in 2D mode, prior starting the application in 3D mode (switching test device to 3D mode);

f) Additionally, at the end of the test the user is asked several additional consistency questions: type of web-browser used, their internet download speed, number of sequences with frame freezing, if their monitor/TV dropped out of 3D mode, location (country) where the test took place and whether they have normal depth acuity.

g) Workers were required to provide some information that provides reasonable proof of them having finished the evaluations through the use of crowdsourcing platform interface pages. One such piece of information is the brand and model type of their 3D display which have to match the brand and model indicated on the test site. For the same reason the user has to submit a picture of the test set-up, showing (in the same picture) the 3D monitor/3D TV set, the 3D glasses used and the test web-site displayed on the screen. This was one final ARM which was implemented on the crowdsourcing platform interface and it ensured that the test web-site is displayed correctly on the user 3D monitor/3D TV set and that all the necessary and right equipment had been used. If the users provided unsatisfactory data for this last verification (for example the picture showed a wrong type of 3D glasses, or did not show the test site loaded on screen) then their results were dismissed, their tasks were rated unsuccessful and they were not paid.

B. Technical challenges

Conceptual requirements to be met by the proposed framework of the *Crowd3D* method are: two-stage design, first 2 sequences used as training, maximum duration of assessment about 20 minutes, additional ARMs listed earlier, optimisation to run on slower computers (it was tested using processor Intel Core2 Duo E8400 @ 3.00GHz).

Traditional test procedures such as ITU-R BT.500-13 [1] and ITU-R BT.2021 [2] can be modified in accordance with the technical and conceptual requirements of the *Crowd3D* method and adapted for use in crowdsourced subjective 3D video quality assessment. In this way a common evaluation ground is established, allowing fair comparison of the results obtained with the *Crowd3D* method with those obtained using traditional laboratory-based subjective 3D video quality assessment methods. In the current work this comparison quantified the degree of agreement between the two sets of grades and led to some conclusions about the reliability of the crowd-based quality grades.

The commercial crowdsourcing platform Microworkers [11] was chosen for conducting the desired crowdsourced 3D video quality assessment. This platform was chosen because of its flexibility in campaign design (usage of the dedicated test server, set up in a laboratory of the research institute Instituto de Telecomunicações in Portugal).

We used multiple design voting scale (gathered grades for quality, depth and comfort scores using 3 voting scales simultaneously, Fig. 1). According to ITU P.915 [4], single or multiple questionnaire is possible. In case of multiple questions it is advisable to consult generally available information from psychology. It would take more time to gather all grades in that case, which may not be suitable for crowdsourced environment. Before each evaluation starts, it has been explained to the observers that: "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". Grades have been collected using 3 sliders in range 0-5, with step 0.1. Voting scale is presented in Fig. 1, in 2D mode, in 3D mode it would be seen as one object (consisting of 3 scales for image, depth and comfort grades). Although it was possible to use discrete 5 point Likert based voting scale, we used continuous grading scale to have the same type of the voting mechanism like in laboratory tests from [13]. Also, by using such type of grading, we could implement safety checks regarding detection of potential cheaters: if users choose more than five default ratings (they do not move the rating sliders for 5 3D video sequences from middle grade, 2.5) or if the average grades of original 3D video sequences was below 1.5. It might be more difficult to choose those boundaries in discrete type of grading, especially using 5 points.

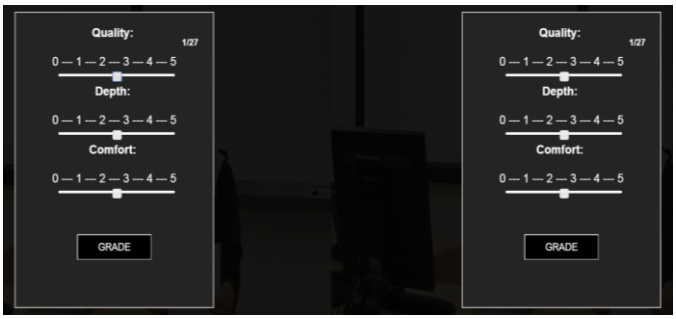


Fig. 1. Voting scale used in 3D crowdsourced assessment

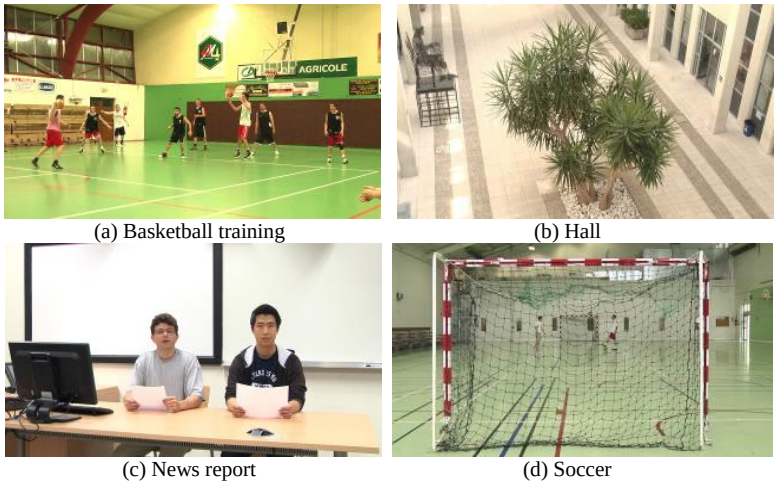
C. Crowd3D system description, implementation and videosequences description

Test campaign of Crowd3D consists of two stages. Both stages of our crowdsourced 3D video quality assessment can be accessed and run online [28], [29], together with instructions given to the observers. Stage one was used in order to assemble a group of pseudo-reliable participants where the screening criterion used was normal depth perception and possession of either a 3D monitor or 3D TV set. Only the participants who passed stage one were allowed to take part in stage two. Stage one screening used only five 3D video sequences. In related work [8] the authors did not test the subjects for vision impairments, instead instructed the workers to use whatever corrective lenses they used in their day-to-day life, during the study. Later in the survey, the subjects were asked if they usually wore corrective lenses and whether they wore the lenses while participating in the study. The ratings given by those subjects who were not wearing their corrective lenses they were otherwise supposed to wear were rejected. In our work workers vision was tested through questionnaires. If the workers stated they do not have normal depth acuity their results were discarded.

Stage two used four original 3D video sequences and 21 corresponding degraded 3D video sequences. This results in 21*4=84 degraded sequences, plus 4 original sequences, equals 88 overall 3D video sequences to give subjective grade.

The 88 sequences (some of them were used in stage one also) 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 (median PSNR=50.9415 dB, median SSIM=0.9959), which show 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 [13].

At the beginning of each evaluation in stage two, 2 additional sequences were used intended to serve as an introduction (reminder) to the observers of the grading system and assessment procedure (those grades were discarded in later analysis). The four original 3D stereo video sequences are available for download from [30]: Basketball training, Hall, News report and Soccer. These four sequences are in full HD stereo format, with 25 fps frame rate and are 16 seconds long. Detailed information about all the sequences used can be found in [22]. The left view of the first frame of each original sequence is presented in Fig. 2 and the spatial and temporal activity indices of those sequences, computed as stated in ITU-T recommendation P.910 [31], are plotted in Fig. 3.



(a) Basketball training (b) Hall

(c) News report (d) Soccer

Fig. 2. First frame, left view, from each of the tested sequences

The activity indices plot shows that the sequences are very diverse in terms of their spatial and temporal characteristics, ensuring that the chosen sequences are a representative sample of the type of contents found in real applications.

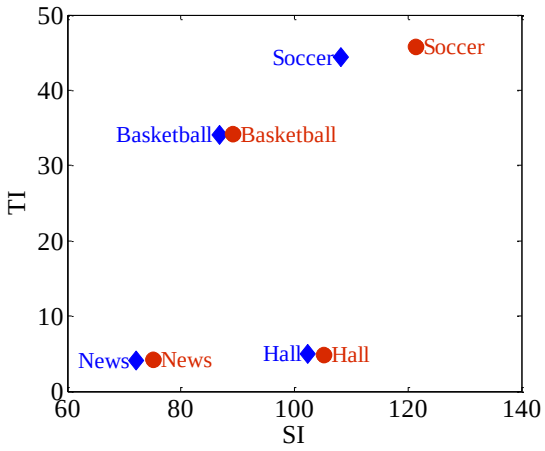


Fig. 3. Spatial versus temporal information: ◊ - left view; ○ - right view

The degradations that were tested in the subjective evaluation are explained in detail in 3DVCL@FER [13]. A palette of 21 degradations was used, including: Compression related degradations (H.264/AVC, HEVC) - 7 types, temporal degradations - 4 types, incorrect camera settings - 5 types, resizing, packet losses, 2D view, compressed 2D view, 2D to 3D conversion. Degradation number '5' from 3DVCL@FER - difference in gamma between left and right view - was not tested here due to the possibility of dropping out from 3D mode in some TV sets (and this cannot be controlled in crowdsource evaluations).

Crowd3D page [29] for the second stage grade collection is also shown in Fig. 4.

(a) Before preloading

(b) After preloading

(c) After 3D video quality assessment

Fig. 4 Crowd3D second stage page: a) starting page, e.g. before preloading; b) after preloading and before start of the 3D video quality assessment; c) after 3D video quality assessment

D. Crowd3D Test Setup and grade collection – stage two

Because of the Google Chrome cache size constraints and the large number of 3D video sequences used in this quality measurement tasks, the content to be evaluated was divided into 4 equal parts, which were preloaded in the Chrome cache. Each part of this data set had a size of approximately 1.3 GB, which is smaller than the maximum allowable size of the Google Chrome cache (1.8 GB). The total time for conducting one part of our crowdsourced 3D video quality assessment was around 15

minutes, not including the preloading time that depends on the user's internet access speed. Furthermore each worker was allowed to participate in several crowdsourced 3D video quality assessment sessions, with a maximum of two sessions per day. In each part, observers evaluated: 4 original sequences, 21 degraded sequences (with different content, e.g. 5-6 degradation types per sequence), plus 2 sequences at the beginning used for introduction (overall 27 sequences per evaluation). So, each session has 27 sequences that last 16 seconds each, overall 432 seconds, plus the time needed for the observers to give their grades (usually under 20 minutes). Two sessions per day were asked so that observers would not get tired from more evaluations. Every time sequences from each part were randomized (only 2 at the beginning, used as an introduction to the observers of the grading system and assessment procedure, always stayed at the beginning). Application also takes information about which of those 4 parts were evaluated, so if an observer gets 2 or more times the same part, only first evaluation of that part was taken into later calculation of the grades, while other evaluations were discarded (although being paid). Average preloading time heavily depends on internet speed of the user. When starting the application, users have information that they will have to download up to 2 GB, and during preloading time, sequence number is being shown which is currently preloading (1 to 27).

IV. Results: Grade Preprocessing and Analysis

A. *Preprocessing of crowdsourced gathered grades*

After all subjective scores were collected, we had successfully finished 220 stage two subjective tests (out of overall 283 sessions that started application). Potential cheaters, as explained earlier, were screened *a priori* with the help of ARMs implemented in the crowdsourcing platform (5 results – observers who did not send correct verification of their equipment) and in the test application itself (13 results – marked as "potential cheater"). In total 18 results were removed with recourse to the ARMs. However, some of the observers evaluated the same part of the 3D video contents several times. In these cases, only the first successfully finished evaluation was used to compile the final grades. After that pruning step, unique evaluations were kept in the records, overall 139 evaluations. On average, each degraded video sequence was graded 34.8 times.

Firstly we calculated Pearson's correlation between each of the 139 observation sessions and average score from all observers for quality, depth, and comfort. Average Pearson's correlation in this case was 0.7609 for quality, 0.5902 for depth and 0.6517 for comfort grades. It can be concluded that highest Pearson's correlation was obtained for quality scores, lower for comfort scores and lowest for depth scores. This shows highest variability in the depth scores, smaller diversity for comfort and smallest for quality scores. As reported later in the discussion, a similar trend was observed comparing crowdsourced and laboratory scores: higher correlation is observed for quality, lower for comfort and depth DMOS/MOS (Difference Mean Opinion Score/ Mean Opinion Score) scores.

A further screening of the observers was performed following the procedure suggested in ITU-R BT.500-13 [1] to discard scores from observers who differed too much from the average value (outliers). This procedure involves several steps described next. As a first step each grade residual (difference between reference and degraded video sequence grade for the same observer) was converted to a *z*-score according to (1).

$$z_{nl} = \frac{d_{nl} - \mu_n}{\sigma_n} \tag{1}$$

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 degraded video sequence *l*, μ_n is the residual mean score from observer *n* and σ_n is the residual standard deviation for the scores from observer *n* (over all degraded sequences *l* graded by 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). A similar procedure is used in [32]. However, DMOS results that skip this step were also analysed. Also, using a similar formula we computed z-scores from raw observers' grades, to be able to calculate normalized MOS scores.

$$z'_{nl} = \frac{r_{nl} - \mu'_n}{\sigma'_n} \tag{2}$$

In (2) z'_{nl} is the z-score of observer *n*, for video sequence *l*, r_{nl} is the raw score of observer *n*, for video sequence *l* (original or degraded), μ'_n is the mean score from observer *n* and σ'_n is the standard deviation for the scores from observer *n* (over all sequences *l* graded by that observer).

For each time window (16 seconds per video sequence) normality of the *z*-scores was tested using kurtosis β, over the span of all *z*-scores for that video sequence. Depending on the kurtosis value, each observer's grade was compared to a multiple of the deviation σ_l from the mean value of each video sequence *l*. Finally, following recommendation ITU-R BT.500-13 [1], the decision of whether or not to consider a score from a given observer an outlier is based on two values, P_n and Q_n, computed according to (3) and which basically count the number of scores that fall on the tails of the probability distribution of the normalized scores.

$$\begin{aligned}
&\forall l \in L \text{ where } L \text{ stands for number of video sequences} \\
&\forall n \in N \text{ where } N \text{ stands for number of observers} \\
&\left.\begin{array}{l}\text{if } z_{nl} \geq \bar{z}_l + 2\cdot\sigma_l \text{ then } P_n = P_n + 1 \\ \text{if } z_{nl} \leq \bar{z}_l - 2\cdot\sigma_l \text{ then } Q_n = Q_n + 1\end{array}\right\} \text{for } 2 \leq \beta \leq 4 \text{ (normal)} \\
&\left.\begin{array}{l}\text{if } z_{nl} \geq \bar{z}_l + \sqrt{20}\cdot\sigma_l \text{ then } P_n = P_n + 1 \\ \text{if } z_{nl} \leq \bar{z}_l - \sqrt{20}\cdot\sigma_l \text{ then } Q_n = Q_n + 1\end{array}\right\} \text{for } \beta \notin [2,4] \text{ (not normal)}
\end{aligned} \tag{3}$$

The P_n and Q_n values represent the number of outlier scores for observer n. These P_n and Q_n values are computed for every observer and if any of them is larger than the respective predetermined threshold P_{thresh} or Q_{thresh} of tested (degraded) video sequences, that observer's data are discarded. For MOS outlier scores calculation, in (3) the z_{nl} from (1) has to be replaced with z'_{nl} from (2). We have defined 4 different cases with different outlier thresholds, listed next as cases 1. – 4.:

Case 1. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=2 with z-scores calculation;
Case 2. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation;
Case 3. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 without z-scores calculation;
Case 4. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=4 with z-scores calculation.

Afterwards, results for every observer were rescaled to the 0-100 range, according to (4) where max(z) and min(z) represent maximum and minimum z-scores over all observers and all video sequences and $dmos_{\text{n,l}}$ / $mos_{\text{n,l}}$ represents the rescaled grade of viewer n and sequence l:

$$\begin{aligned}
dmos_{n,l} &= \frac{100}{\max(z) - \min(z)} \cdot (z_{n,l} - \min(z)) \\
mos_{n,l} &= \frac{100}{\max(z') - \min(z')} \cdot (z'_{n,l} - \min(z'))
\end{aligned} \tag{4}$$

At the end, an average DMOS(l) grade was calculated for each of the distorted video sequences as the arithmetic mean of all grades for that sequence. In every evaluation session the observer graded videos covering all types of degradations so no degradation specific bias occurred. Consequently, there was no need for further realignment of the DMOS scores.

Fig. 5 shows different factors from the crowdsourced test environment that can influence the final grades: age, gender, device type (TV, monitor, laptop) and glasses type (active or passive). To understand the impact of some of these factors on the scores, an analysis presented later was done according to observer gender - males, observers who used TV sets, observers with ages from 20 to 35 years and observers who used active glasses display. Due to lower number of observers in other groups - female observers, observers who used monitor, laptop etc., results are not shown because they could be unreliable. For that analysis, we always used *DMOS/MOS* scores with P_{thresh}=Q_{thresh}=3, with z-scores calculation, subset of case 2. So, additionally we tested another 4 different cases (cases 5. –8.):

Case 5. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation - Males only;
Case 6. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation - TV sets only;
Case 7. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation - 20-35 years observers only;
Case 8. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation - Active glasses display.

Finally, we also checked influence of previously discussed ARMs on final DMOS/MOS score correlation by defining two new cases (cases 9.-10.) which include scores rejected when the ARM are enforced, and proceeding as in the previous cases analyses. The new cases are defined as follows:

Case 9. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation, together with 13 potential cheater sessions and 5 session who did not pass final verification test (12 "false" sessions added overall – because some are overlapping and some were also double sessions), which gives 151 sessions overall;
Case 10. *DMOS/MOS* scores, P_{thresh}=Q_{thresh}=3 with z-scores calculation, together with false sessions described earlier and double sessions, added 81 sessions, which gives 220 sessions overall;

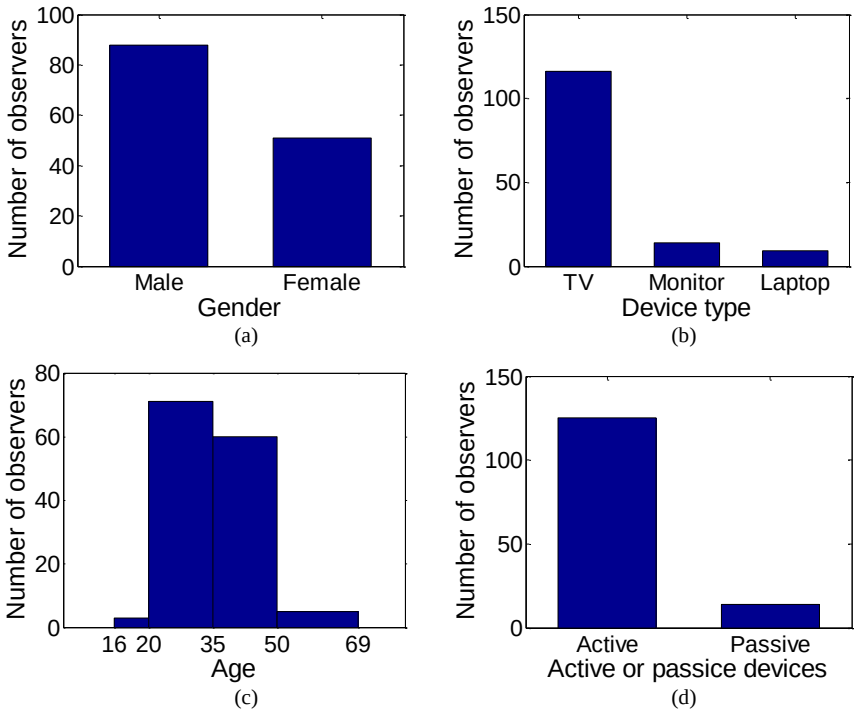


Fig. 5. Different factors from crowdsourced test: (a) Gender, (b) Device type, (c) Age, (d) Active/passive devices

The number of discarded observers for quality, depth and comfort scores, as well as the total number of discarded observers for the previously described cases 1-10 are shown in Table I. It can be seen that z-score calculation rearranges score spans from different observers into more similar range, resulting in less outliers: the P_{thresh}=Q_{thresh}=3, with z-scores calculation case has less number of discarded observers than the case for P_{thresh}=Q_{thresh}=3, without z-scores calculation. Also, as expected, with higher values for the thresholds P_{thresh} and Q_{thresh}, the total number of discarded observers is smaller.

TABLE I NUMBER OF DISCARDED OBSERVERS FOR DIFFERENT VALUES P_{THRESH} AND Q_{THRESH}

		Discarded observers for			Overall number of discarded observers
		Quality	Depth	Comfort	
DMOS scores	1. Crowdsourced test, P_{thresh}=Q_{thresh}=2, with z-scores	17	13	17	34
	2. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores	1	2	1	4
	3. Crowdsourced test, P_{thresh}=Q_{thresh}=3, skip z-scores	7	12	11	21
	4. Crowdsourced test, P_{thresh}=Q_{thresh}=4, with z-scores	0	0	0	0
	5. Males only test, P_{thresh}=Q_{thresh}=3, with z-scores	2	0	0	2
	6. TV sets only, P_{thresh}=Q_{thresh}=3, with z-scores	2	1	2	5
	7. 20-35 years observers only, P_{thresh}=Q_{thresh}=3, with z-scores	2	0	0	2
	8. Active glasses display, P_{thresh}=Q_{thresh}=3, with z-scores	3	1	1	5
	9. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with false results	3	0	0	3
	10. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with added results from false and double results	2	3	2	6
MOS scores	1. Crowdsourced test, P_{thresh}=Q_{thresh}=2, with z-scores	19	18	14	39
	2. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores	3	3	5	10
	3. Crowdsourced test, P_{thresh}=Q_{thresh}=3, skip z-scores	8	15	11	25
	4. Crowdsourced test, P_{thresh}=Q_{thresh}=4, with z-scores	1	0	1	2
	5. Males only test, P_{thresh}=Q_{thresh}=3, with z-scores	2	1	2	5
	6. TV sets only, P_{thresh}=Q_{thresh}=3, with z-scores	2	1	2	5
	7. 20-35 years observers only, P_{thresh}=Q_{thresh}=3, with z-scores	3	0	2	4
	8. Active glasses display, P_{thresh}=Q_{thresh}=3, with z-scores	5	3	2	8
	9. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with false results	5	3	7	12
	10. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with added results from false and double results	9	3	5	16

B. Comparative assessment of laboratory and crowd-based grades

An important objective of this work is understanding if the crowdsource based quality evaluation results are similar to results of studies performed in more controlled conditions in a laboratory. To do this analysis we started by applying a nonlinear regression function to the two sets of data, to compensate for the fact that the laboratory DMOS/MOS results were obtained using more video sequences than the crowdsource study and so the two sets of DMOS/MOS values (laboratory and crowdsource) did not have the same span. The laboratory DMOS/MOS scores were preprocessed by removing all raw observers' evaluations that did not grade sequences also used in the crowdsourced study. Then the procedure used to compute the DMOS/MOS values in the crowdsourced case (using (1), (2), (3) and (4)), was applied to the filtered laboratory scores, with P_{thresh}=Q_{thresh}=3 and z-scores calculation, i.e. case 2 from above. Choosing only 4 sequence types *a priori* is possible because in laboratory evaluations, observers watched either 4 sequence types (original and degraded) present in crowdsourced experiment, or the other 4 sequence types and no observer was presented a mix of these two 4 types sets. In the laboratory evaluations, 15-20 grades per each degraded video sequence were collected (after 1 outlier removal) for both MOS and DMOS scores.

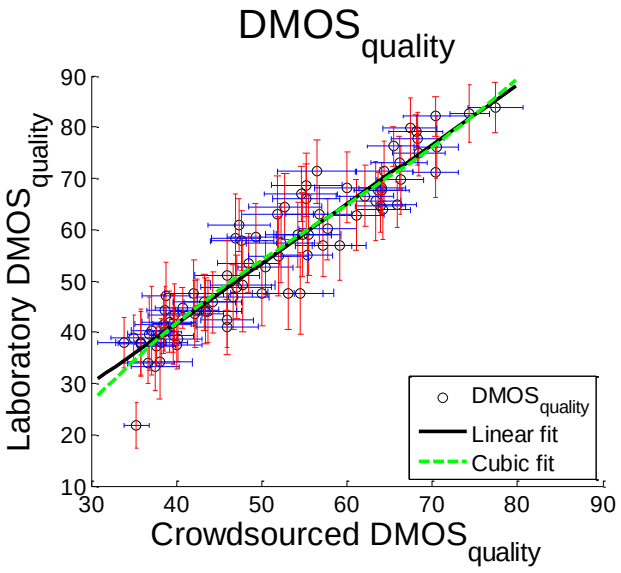
Then three different possibilities were considered to obtain an analytical description of the data, no fit, linear and cubic polynomial fit (best fit in a least-squares sense), like in [14]. These fit alternatives formulations are as listed in (5).

$$\begin{aligned} Q_{No-fit}(z) &= z \\ Q_{linear}(z) &= b_1 \cdot z + b_2 \\ Q_{cubic}(z) &= b_1 \cdot z^3 + b_2 \cdot z^2 + b_3 \cdot z + b_4 \end{aligned} \tag{5}$$

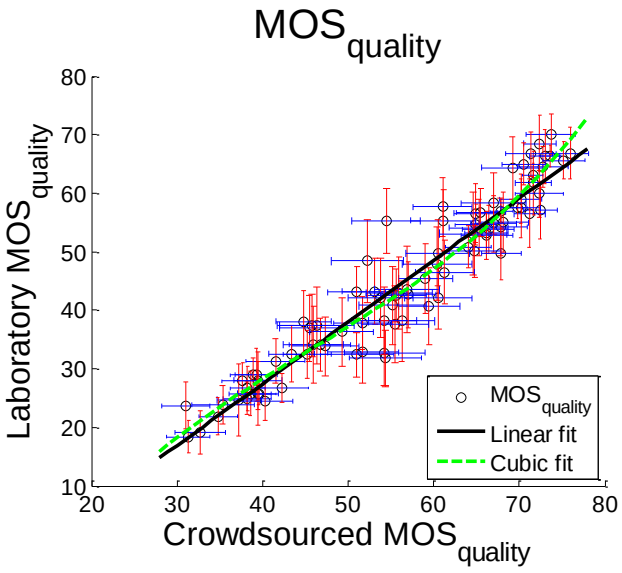
Fig. 6 shows the linear and quadratic fitted functions for the DMOS/MOS data for the three quality dimensions, "Quality", "Depth" and "Comfort", when using P_{thresh}=Q_{thresh}=3, and z-scores calculation (previously described case 2) . Table II lists the values of the parameters of the fitted models.

TABLE II PARAMETERS USED TO FIT BETWEEN CROWDSOURCED AND LABORATORY DMOS/MOS SCORES AND 95% CONFIDENCE INTERVAL

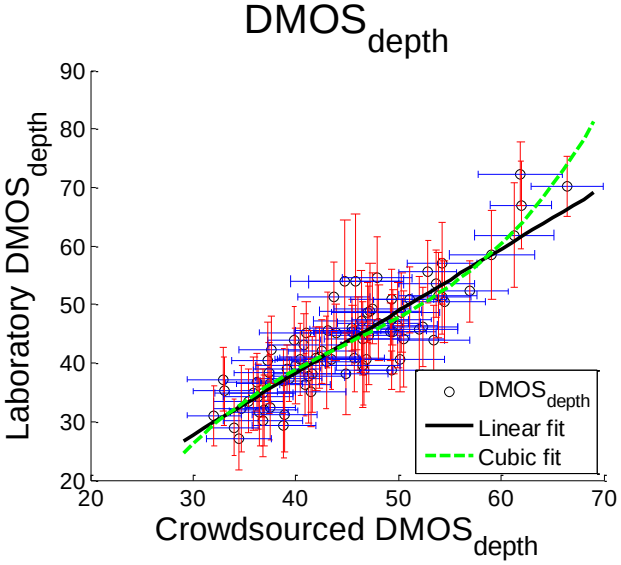
Score type		b_1 (95% CI)	b_2 (95% CI)	b_3 (95% CI)	b_4 (95% CI)
DMOS scores, Linear fit	Quality	1.1655 (±0.0774)	-4.9600 (±4.067)	-	-
	Depth	1.0589 (±0.1070)	-4.0371 (±4.8686)	-	-
	Comfort	0.8246 (±0.0775)	11.0851 (±3.5498)	-	-
DMOS scores, Cubic fit	Quality	0.0003032 (±0.0006454)	-0.0527 (±0.1053)	4.1275 (±5.5904)	-58.4010 (±96.4503)
	Depth	0.0012402 (±0.0012203)	-0.1676 (±0.1771)	8.4152 (±8.4095)	-108.8645 (±130.6183)
	Comfort	0.0002776 (±0.0007114)	-0.0406 (±0.0983)	2.7321 (±4.3999)	-17.3842 (±63.3112)
MOS scores, Linear fit	Quality	1.0549 (±0.0577)	-14.7542 (±3.3281)	-	-
	Depth	1.1159 (±0.0847)	-14.1282 (±5.1510)	-	-
	Comfort	0.8745 (±0.0738)	3.7410 (±4.2335)	-	-
MOS scores, Cubic fit	Quality	0.0003274 (±0.0004323)	-0.0449 (±0.0701)	2.9344 (±3.6906)	-38.2622 (±62.7899)
	Depth	0.0012677 (±0.0007044)	-0.2179 (±0.1212)	13.3661 (±6.8537)	-238.9810 (±127.3323)
	Comfort	0.0006516 (±0.0007730)	-0.1067 (±0.1312)	6.5579 (±7.2953)	-94.6653 (±132.8156)


DMOS quality
Laboratory DMOS quality
Crowdsourced DMOS quality
DMOS quality
Linear fit
Cubic fit

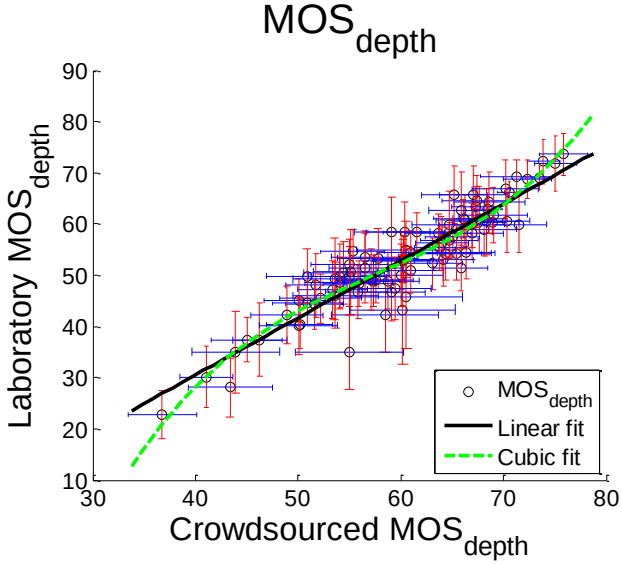
(a)


MOS quality
Laboratory MOS quality
Crowdsourced MOS quality
MOS quality
Linear fit
Cubic fit

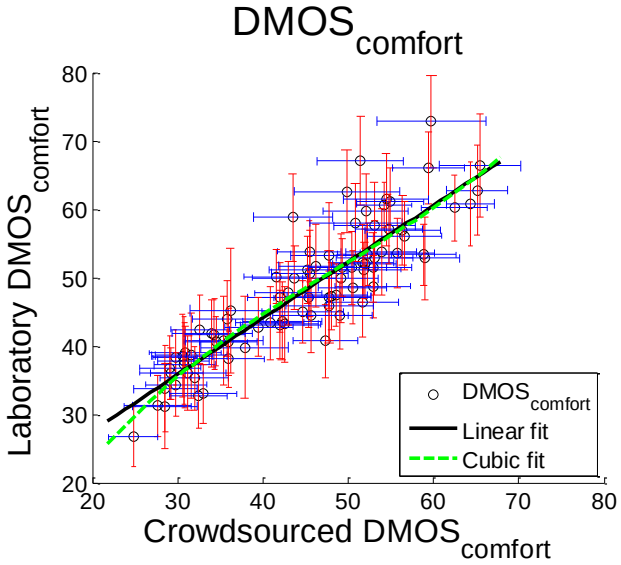
(b)

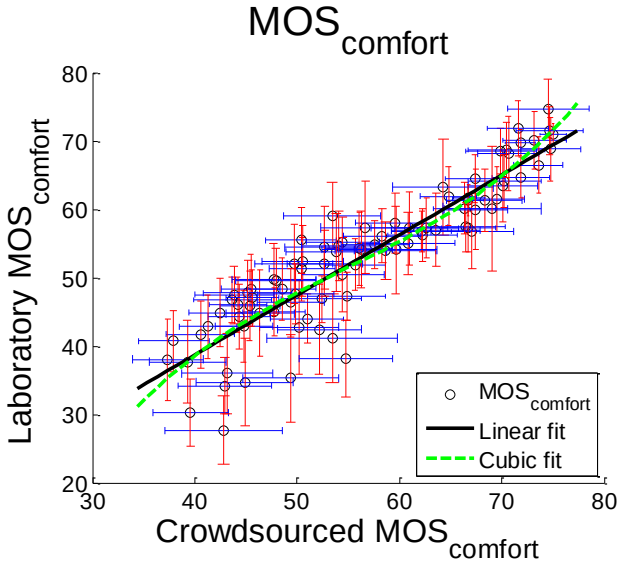

DMOS depth
Laboratory DMOS depth
Crowdsourced DMOS depth
DMOS depth
Linear fit
Cubic fit

(c)


MOS depth
Laboratory MOS depth
Crowdsourced MOS depth
MOS depth
Linear fit
Cubic fit

(d)


DMOS comfort
Laboratory DMOS comfort
Crowdsourced DMOS comfort
DMOS comfort
Linear fit
Cubic fit


MOS comfort
Laboratory MOS comfort
Crowdsourced MOS comfort
MOS comfort
Linear fit
Cubic fit

(e) (f)

Fig. 6. Crowdsourced and laboratory results comparison, together with 95% CI in related direction, for: (a) DMOS quality, (b) MOS quality, (c) DMOS depth, (d) MOS depth, (e) DMOS comfort, (c) MOS comfort

The DMOS/MOS scores for the original sequences are shown in Table III, together with 95% CI. Because DMOS scores are the same for all original sequences, they were not used for later comparison with laboratory test in DMOS cases comparison (as different number of original sequences would change correlation, RMSE etc.). In the case of MOS scores, the grades for the original video sequences have been included in the correlation calculations.

TABLE III DMOS/MOS SCORES FOR ORIGINAL SEQUENCES

Score type		Original sequence "Basketball training"		Original sequence "Hall"		Original sequence "News report"		Original sequence "Soccer"	
		crowdsourced (95% CI)	laboratory (95% CI)	crowdsourced (95% CI)	laboratory (95% CI)	crowdsourced (95% CI)	laboratory (95% CI)	crowdsourced (95% CI)	laboratory (95% CI)
DMOS scores	Quality	36.9556 (±0.8610)	37.4653 (±1.3208)	36.9556 (±0.8610)	37.4653 (±1.3208)	36.9556 (±0.8610)	37.4653 (±1.3208)	36.9556 (±0.8610)	37.4653 (±1.3208)
	Depth	34.6221 (±0.8287)	31.9060 (±1.1310)	34.6221 (±0.8287)	31.9060 (±1.1310)	34.6221 (±0.8287)	31.9060 (±1.1310)	34.6221 (±0.8287)	31.9060 (±1.1310)
	Comfort	30.1558 (±1.0025)	35.0346 (±1.3394)	30.1558 (±1.0025)	35.0346 (±1.3394)	30.1558 (±1.0025)	35.0346 (±1.3394)	30.1558 (±1.0025)	35.0346 (±1.3394)
MOS scores	Quality	67.8601 (±1.3264)	56.8852 (±2.028)	71.5962 (±1.1587)	63.1418 (±2.0135)	69.8443 (±1.1304)	57.5043 (±1.8844)	73.6031 (±1.2279)	66.4322 (±1.8431)
	Depth	68.4990 (±1.3786)	62.8564 (±2.3292)	70.6167 (±1.2479)	66.2525 (±2.0188)	68.9717 (±1.2267)	60.3801 (±2.3382)	73.4790 (±1.3380)	68.9923 (±2.4163)
	Comfort	66.7045 (±1.5704)	61.2939 (±2.3092)	71.8040 (±1.4988)	69.7488 (±1.9833)	66.7560 (±1.6014)	61.6603 (±1.9902)	74.9985 (±1.2946)	71.0297 (±1.5970)

The DMOS/MOS scores for quality, depth and comfort were compared with DMOS/MOS scores from laboratory DMOS/MOS scores in 3DVCL@FER using Pearson's and Spearman's correlation, RMSE (Root Mean Square, after nonlinear regression) and OR (Outlier Ratio, after nonlinear regression). To measure the agreement between the two sets of DMOS/MOS values, two figures of merit were used; Root Mean Square Error (RMSE) and Outlier Ratio (OR). RMSE was calculated according to (6)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(fit((D)MOS_{crowd}(i)) - (D)MOS_{lab}(i))^2} \tag{6}$$

where N represents number of tested video sequences (in our case 84 degraded video sequences for DMOS or 88 for MOS scores), $fit((D)MOS_{crowd}(i))$ is the fitted *(D)MOS* score of the *i*-th video sequence in crowdsourced test (using (5)) and $(D)MOS_{lab}(i)$ is *(D)MOS* laboratory obtained score of *i*-th video sequence.

The other indicator, OR, was calculated as the number of video sequences that fall outside the 95% confidence interval calculated from DMOS/MOS laboratory tests where the confidence interval is computed by (7)

$$CI = t(M-1)\cdot\frac{std(scores)}{\sqrt{M}} \tag{7}$$

where t(M) is the critical value of Student's t distribution with M-1 degrees of freedom (M is number of times that the same video sequence has been graded) for 95% probability and std(scores) is the standard deviation of the grades of the same video sequence. A score was considered to be an outlier if (8) held

$$\left| fit((D)MOS_{crowd}(i)) - (D)MOS_{lab}(i) \right| > CI_{crowd}(i) + CI_{lab}(i) \tag{8}$$

and OR was calculated as number of outliers divided by the number of tested video sequences (in our case 84 degraded video sequences for DMOS or 88 for MOS scores).

Separate analysis of the crowd-based grades was performed taking into consideration the previously described cases 1-10. Pearson's and Spearman's inter-correlation (without any fitting) between DMOS/MOS scores for quality, depth and comfort, for results from laboratory scores, crowdsourced (with different P_{thresh} and Q_{thresh} values), males only, TV sets only, 20-35 years observers only, active glasses display, results with false observers' scores, results with false and double observers' scores, are presented in Table IV.

TABLE IV PEARSON'S AND SPEARMAN'S INTER-CORRELATION BETWEEN DMOS/MOS SCORES FOR QUALITY, DEPTH AND COMFORT

		Short label for each case	Pearson's correlation between scores for			Spearman's correlation between scores for		
			Quality & Depth	Depth & Comfort	Comfort & Quality	Quality & Depth	Depth & Comfort	Comfort & Quality
DMOS scores	0. Laboratory test, P_{thresh}=Q_{thresh}=3, with z-scores	0. Lab-DMOS	0.6317	0.5006	0.6596	0.6109	0.5685	0.7214
	1. Crowdsourced test, P_{thresh}=Q_{thresh}=2, with z-scores	1. PQ2	0.8320	0.7589	0.8194	0.8039	0.7901	0.8198
	2. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores	2. PQ3	0.8355	0.7577	0.8377	0.8100	0.7810	0.8357
	3. Crowdsourced test, P_{thresh}=Q_{thresh}=3, skip z-scores	3. PQZ3	0.8318	0.7231	0.8186	0.8142	0.7543	0.8345
	4. Crowdsourced test, P_{thresh}=Q_{thresh}=4, with z-scores	4. PQ4	0.8402	0.7564	0.8376	0.8120	0.7779	0.8392
	5. Males only test, P_{thresh}=Q_{thresh}=3, with z-scores	5. Males	0.8186	0.6850	0.7755	0.7898	0.7235	0.7959
	6. TV sets only, P_{thresh}=Q_{thresh}=3, with z-scores	6. TV sets	0.8097	0.7162	0.8076	0.7858	0.7484	0.8003
	7. 20-35 years observers only, P_{thresh}=Q_{thresh}=3, with z-scores	7. 20-35 y	0.8293	0.7192	0.7654	0.8093	0.7423	0.7446
	8. Active glasses display, P_{thresh}=Q_{thresh}=3, with z-scores	8. AGD	0.8353	0.7714	0.8694	0.8210	0.7936	0.8650
	9. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with false results	9. ARM1	0.8460	0.7650	0.8402	0.8167	0.7824	0.8452
	10. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with added results from false and double results	10. ARM2	0.8963	0.8095	0.8336	0.8584	0.8267	0.8602
MOS scores	0. Laboratory test, P_{thresh}=Q_{thresh}=3, with z-scores	0. Lab-MOS	0.6442	0.5480	0.6735	0.6228	0.6105	0.7092
	1. Crowdsourced test, P_{thresh}=Q_{thresh}=2, with z-scores	1. PQ2-M	0.8163	0.7736	0.8480	0.8125	0.8038	0.8511
	2. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores	2. PQ3-M	0.8305	0.7988	0.8607	0.8205	0.8251	0.8640
	3. Crowdsourced test, P_{thresh}=Q_{thresh}=3, skip z-scores	3.PQZ3-M	0.7954	0.7701	0.8367	0.7821	0.7997	0.8490
	4. Crowdsourced test, P_{thresh}=Q_{thresh}=4, with z-scores	4. PQ4-M	0.8374	0.7954	0.8503	0.8263	0.8284	0.8530
	5. Males only test, P_{thresh}=Q_{thresh}=3, with z-scores	5. Males-M	0.8065	0.7380	0.7899	0.7783	0.7753	0.8239
	6. TV sets only, P_{thresh}=Q_{thresh}=3, with z-scores	6. TVsets-M	0.8099	0.7638	0.8339	0.8002	0.7833	0.8280
	7. 20-35 years observers only, P_{thresh}=Q_{thresh}=3, with z-scores	7. 20-35y-M	0.8352	0.7570	0.7809	0.8319	0.7904	0.7844
	8. Active glasses display, P_{thresh}=Q_{thresh}=3, with z-scores	8. AGD-M	0.8312	0.8033	0.8728	0.8188	0.8161	0.8730
	9. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with false results	9. ARM1-M	0.8418	0.7997	0.8598	0.8332	0.8346	0.8593
	10. Crowdsourced test, P_{thresh}=Q_{thresh}=3, with z-scores, with added results from false and double results	10. ARM2-M	0.8910	0.8324	0.8484	0.8683	0.8633	0.8753

The DMOS and MOS scores for quality, depth and comfort collected in all previously defined evaluations/groups were compared with DMOS and MOS scores from the laboratory evaluations using Pearson's and Spearman's correlation (together with confidence interval), RMSE (Root Mean Square Error) and OR (Outlier Ratio) and presented in Table V and Table VI respectively. The best results are highlighted in bold. In the case of MOS calculation, we had overall 88 MOS scores, 4 more than DMOS, because the original video sequences were also taken into account. Confidence intervals for Pearson's and Spearman's correlations have been calculated using Fisher's transform (9). It has to be noted that overlapping CIs do not necessarily mean that correlations are statistically similar.

$$lower\_bound = \tanh\left(\operatorname{atanh}(\rho) - z_{CI}\cdot\frac{1}{\sqrt{N_{videosequences}-3}}\right)$$
$$upper\_bound = \tanh\left(\operatorname{atanh}(\rho) + z_{CI}\cdot\frac{1}{\sqrt{N_{videosequences}-3}}\right) \quad (9)$$

ρ – Pearson's or Spearman's correlation

$z_{CI} = 1.9600$ for CI=95%

TABLE V PEARSON'S CORRELATION, SPEARMAN'S CORRELATION, RMSE AND OR BETWEEN **DMOS** SCORES FROM DIFFERENT SUBSETS OF CROWDSOURCED EVALUATION AND LABORATORY TEST; LABELS FOR EACH CASE ARE DESCRIBED IN TABLE IV

	min number of obs.	max number of obs.	Pearson's correlation between scores for			Spearman's correlation between scores for			RMSE between scores for			OR between scores for		
			quality (95% CI)	depth (95% CI)	comfort (95% CI)	quality (95% CI)	depth (95% CI)	comfort (95% CI)	quality	depth	comfort	quality	depth	comfort
1. PQ2 – Cubic fit	24	29	0.9356 (0.9021 - 0.9578)	**0.8943** (0.8412 - 0.9303)	0.8930 (0.8393 - 0.9295)	0.9391 (0.9074 - 0.9602)	0.8465 (0.7723 - 0.8980)	0.9075 (0.8605 - 0.9392)	4.9856	**4.0186**	4.2820	0.0119	0.0119	0.0238
2. PQ3 – No fit	31	37	0.9405 (0.9095 - 0.9611)	0.8761 (0.8148 - 0.9181)	0.8902 (0.8353 - 0.9276)	0.9455 (0.9170 - 0.9644)	0.8528 (0.7813 - 0.9022)	0.9051 (0.8570 - 0.9376)	6.2493	4.5689	5.7135	0.1667	0.0476	0.0952
2. PQ3 – Linear fit	31	37	0.9405 (0.9095 - 0.9611)	0.8761 (0.8148 - 0.9181)	0.8902 (0.8353 - 0.9276)	0.9455 (0.9170 - 0.9644)	0.8528 (0.7813 - 0.9022)	0.9051 (0.8570 - 0.9376)	4.7968	4.3295	4.3338	0.0357	0.0238	0.0357
2. PQ3 – Cubic fit	31	37	0.9414 (0.9108 - 0.9617)	0.8856 (0.8285 - 0.9244)	0.8912 (0.8367 - 0.9283)	0.9455 (0.9170 - 0.9644)	0.8528 (0.7813 - 0.9022)	0.9051 (0.8570 - 0.9376)	4.7614	4.1719	4.3151	0.0238	0.0119	0.0357
3. PQZ3 – Cubic fit	27	32	0.9242 (0.8852 - 0.9503)	0.8714 (0.8080 - 0.9149)	0.8735 (0.8110 - 0.9163)	0.9284 (0.8915 - 0.9531)	0.8375 (0.7594 - 0.8917)	0.8878 (0.8316 - 0.9259)	5.3915	4.4057	4.6322	0.0833	0.0238	0.0357
4. PQ4 – Cubic fit	31	38	0.9408 (0.9100 - 0.9613)	0.8866 (0.8299 - 0.9251)	0.8913 (0.8368 - 0.9283)	0.9442 (0.9150 - 0.9635)	**0.8568** (0.7870 - 0.9050)	0.9059 (0.8581 - 0.9381)	4.7834	4.1543	4.3145	0.0476	0.0119	0.0357
5. Males – Cubic fit	18	25	0.9372 (0.9046 - 0.9589)	0.8800 (0.8204 - 0.9207)	0.8747 (0.8127 - 0.9171)	0.9387 (0.9069 - 0.9599)	0.8364 (0.7579 - 0.8910)	0.8874 (0.8312 - 0.9257)	4.9227	4.2659	4.6112	0.0119	0	0.0119
6. TV sets – Cubic fit	23	32	**0.9456** (0.9172 - 0.9645)	0.8757 (0.8142 - 0.9178)	0.9132 (0.8690 - 0.9430)	**0.9470** (0.9193 - 0.9654)	0.8360 (0.7573 - 0.8907)	**0.9229** (0.8833 - 0.9494)	**4.5919**	4.3361	3.8765	0.0238	0.0119	0
7- 20-35 y – Cubic fit	17	21	0.9358 (0.9025 - 0.9580)	0.8728 (0.8099 - 0.9158)	**0.9135** (0.8693 - 0.9432)	0.9349 (0.9011 - 0.9574)	0.8336 (0.7539 - 0.8891)	0.9124 (0.8678 - 0.9425)	4.9750	4.3838	**3.8716**	0.0119	0.0238	0.0119
8. AGD – Cubic fit	27	34	0.9390 (0.9073 - 0.9601)	0.8861 (0.8292 - 0.9248)	0.8829 (0.8246 - 0.9227)	0.9431 (0.9134 - 0.9628)	0.8499 (0.7771 - 0.9003)	0.9048 (0.8566 - 0.9374)	4.8534	4.1629	4.4671	0.0357	0.0119	0.0357
9. ARM1 – Cubic fit	33	40	0.9376 (0.9052 - 0.9592)	0.8807 (0.8215 - 0.9212)	0.8785 (0.8182 - 0.9197)	0.9419 (0.9116 - 0.9620)	0.8543 (0.7834 - 0.9033)	0.8913 (0.8369 - 0.9283)	4.9075	4.2534	4.5451	0.0357	0.0119	0.0595
10. ARM2 – Cubic fit	48	59	0.9272 (0.8897 - 0.9523)	0.8929 (0.8392 - 0.9294)	0.8628 (0.7955 - 0.9090)	0.9313 (0.8957 - 0.9550)	0.8489 (0.7756 - 0.8995)	0.8755 (0.8139 - 0.9176)	5.2861	4.0430	4.8107	0.0595	0.0357	0.0833

TABLE VI PEARSON'S CORRELATION, SPEARMAN'S CORRELATION, RMSE AND OR BETWEEN **MOS** SCORES FROM DIFFERENT SUBSETS OF CROWDSOURCED EVALUATION AND LABORATORY TEST; LABELS FOR EACH CASE ARE DESCRIBED IN TABLE IV

	min number of obs.	max number of obs.	Pearson's correlation between scores for			Spearman's correlation between scores for			RMSE between scores for			OR between scores for		
			quality (95% CI)	depth (95% CI)	comfort (95% CI)	quality (95% CI)	depth (95% CI)	comfort (95% CI)	quality	depth	comfort	quality	depth	comfort
1. PQ2-M – Cubic fit	23	27	0.9556 (0.9328 - 0.9707)	**0.9379** (0.9066 - 0.9590)	0.9094 (0.8646 - 0.9398)	0.9494 (0.9236 - 0.9666)	**0.9120** (0.8685 - 0.9416)	0.9208 (0.8814 - 0.9475)	4.0998	**3.4155**	4.2880	0.0227	0	0.0341
2. PQ3-M – No fit	30	35	0.9564 (0.9341 - 0.9713)	0.9209 (0.8815 - 0.9476)	0.9048 (0.8580 - 0.9368)	0.9559 (0.9333 - 0.9710)	0.9063 (0.8601 - 0.9377)	0.9200 (0.8802 - 0.9470)	12.3714	8.1638	5.6735	0.8409	0.3750	0.1705
2. PQ3-M – Linear fit	30	35	0.9564 (0.9341 - 0.9713)	0.9209 (0.8815 - 0.9476)	0.9048 (0.8580 - 0.9368)	0.9559 (0.9333 - 0.9710)	0.9063 (0.8601 - 0.9377)	0.9200 (0.8802 - 0.9470)	4.0607	3.8384	4.3894	0.0682	0.0114	0.0341
2. PQ3-M – Cubic fit	30	35	0.9607 (0.9405 - 0.9741)	0.9288 (0.8931 - 0.9529)	0.9078 (0.8624 - 0.9388)	0.9559 (0.9333 - 0.9710)	0.9063 (0.8601 - 0.9377)	0.9200 (0.8802 - 0.9470)	3.8613	3.6487	4.3232	0.0455	0.0114	0.0341
3. PQZ3-M – Cubic fit	26	31	0.9537 (0.9300 - 0.9695)	0.9123 (0.8689 - 0.9418)	0.9162 (0.8746 - 0.9444)	0.9466 (0.9194 - 0.9648)	0.8848 (0.8289 - 0.9231)	0.9254 (0.8881 - 0.9506)	4.1853	4.0326	4.1309	0	0	0.0114
4. PQ4-M – Cubic fit	31	37	**0.9645** (0.9462 - 0.9766)	0.9297 (0.8944 - 0.9535)	0.9116 (0.8679 - 0.9413)	**0.9600** (0.9395 - 0.9737)	0.9012 (0.8527 - 0.9343)	0.9260 (0.8890 - 0.9510)	**3.6745**	3.6271	4.2376	0.0227	0.0227	0.0341
5. Males-M – Cubic fit	18	23	0.9583 (0.9369 - 0.9726)	0.9139 (0.8712 - 0.9429)	0.8902 (0.8367 - 0.9268)	0.9531 (0.9291 - 0.9691)	0.8844 (0.8284 - 0.9229)	0.9050 (0.8583 - 0.9369)	3.9740	3.9979	4.6971	0.0114	0.0114	0.0341
6. TVsets-M – Cubic fit	24	32	0.9614 (0.9416 - 0.9746)	0.9317 (0.8973 - 0.9548)	**0.9330** (0.8992 - 0.9557)	0.9539 (0.9303 - 0.9696)	0.9105 (0.8663 - 0.9406)	**0.9401** (0.9097 - 0.9604)	3.8261	3.5773	**3.7111**	0.0341	0.0114	0.0114
7. 20-35y-M– Cubic fit	16	22	0.9522 (0.9277 - 0.9685)	0.9114 (0.8676 - 0.9412)	0.9126 (0.8693 - 0.9420)	0.9461 (0.9187 - 0.9644)	0.8802 (0.8224 - 0.9201)	0.9122 (0.8688 - 0.9417)	4.2509	4.0517	4.2146	0.0227	0.0114	0.0114
8. AGD-M – Cubic fit	27	33	0.9619 (0.9423 - 0.9749)	0.9239 (0.8859 - 0.9496)	0.9027 (0.8549 - 0.9353)	0.9574 (0.9356 - 0.9720)	0.8891 (0.8352 - 0.9261)	0.9206 (0.8810 - 0.9474)	3.8012	3.7679	4.4360	0.0455	0.0114	0.0568
9. ARM1-M– Cubic fit	32	38	0.9633 (0.9444 - 0.9759)	0.9293 (0.8938 - 0.9532)	0.9100 (0.8655 - 0.9402)	0.9565 (0.9342 - 0.9713)	0.9037 (0.8564 - 0.9360)	0.9222 (0.8834 - 0.9484)	3.7333	3.6372	4.2754	0.0341	0.0114	0.0341
10. ARM2-M – Cubic fit	45	57	0.9568 (0.9347 - 0.9715)	0.9142 (0.8716 - 0.9430)	0.8854 (0.8298 - 0.9236)	0.9525 (0.9283 - 0.9687)	0.8834 (0.8270 - 0.9222)	0.9001 (0.8512 - 0.9336)	4.0441	3.9918	4.7922	0.0682	0.0568	0.1023

C. ANOVA statistical test and error classification

To determine whether the difference between two sets of scores corresponding to the same stereo pair evaluated in crowdsourced test and laboratory test is statistically significant, a multiple comparison test based on ANOVA was performed at a 5% significance level on the scores for quality, depth and comfort (using linear and cubic regression of scores from crowdsourced test, $P_{thresh}=Q_{thresh}=3$, with z-scores calculation, previously described case 2). The results are presented in Table VII. Results show that number of video sequences with unequal mean is the same for DMOS quality, depth and comfort scores, (for case 2, $P_{thresh}=Q_{thresh}=3$, with z-scores, cubic fit). For MOS scores, depth and comfort scores have lower number of video sequences with unequal mean, comparing with quality MOS scores (for case 2, $P_{thresh}=Q_{thresh}=3$, with z-scores, cubic fit).

TABLE VII ANOVA STATISTICAL TEST AND ERROR CLASSIFICATION FOR DMOS/MOS QUALITY, DEPTH AND COMFORT SCORES

		Overall number of video sequences	Number of video sequences with unequal mean for			Percentage of video sequences with equal mean		
			Quality	Depth	Comfort	Quality	Depth	Comfort
DMOS scores	Crowdsourced test, $P_{thresh}=Q_{thresh}=3$, with z-scores, linear fit	84	4	5	7	95.24%	94.05%	91.67%
	Crowdsourced test, $P_{thresh}=Q_{thresh}=3$, with z-scores, cubic fit	84	4	4	4	95.24%	95.24%	95.24%
MOS scores	Crowdsourced test, $P_{thresh}=Q_{thresh}=3$, with z-scores, linear fit	88	9	3	9	89.77%	96.59%	89.77%
	Crowdsourced test, $P_{thresh}=Q_{thresh}=3$, with z-scores, cubic fit	88	6	3	3	93.18%	96.59%	96.59%

In recommendation ITU-T J.149 [33], it is suggested computing the classification errors and use them to evaluate the performance of an objective metric. In this context a classification error is made when the objective metric and subjective test lead to different conclusions (regarding statistical difference of the scores) on a pair of video sequences, *i* and *j*. In the work [34], this methodology was extended to the case of comparison of a pair of subjective tests of 3D video sequences, *i* and *j*, corresponding to quality grades *(D)MOS(i)* and *(D)MOS(j)*, of 3D content on different monitors in subjective laboratory tests. Similarly, we used those classification errors to compare the laboratory evaluations with the crowdsourced evaluations. DMOS/MOS scores from video sequences *i* and *j* were compared using (10), analogously to (8) where we used same CI as defined in (7):

$$\begin{aligned} &\left| fit((D)MOS_{crowd}(i)) - fit((D)MOS_{crowd}(j)) \right| > CI_{crowd}(i) + CI_{cword}(j) \\ &\&\& \\ &\left| (D)MOS_{crowd}lab(i) - (D)MOS_{lab}(j) \right| > CI_{lab}(i) + CI_{lab}(j) \end{aligned} \tag{10}$$

Borrowing the notation introduced in [34], three types of classification errors are defined:

a) False Tie, the least offensive error. It happens when the laboratory evaluation says that DMOS/MOS scores of sequences *i* and *j* are different (their CIs do not overlap) whereas the evaluation in crowdsourced test says that they are identical (their CIs overlap),

b) False Differentiation: it happens when the evaluation in laboratory test says that DMOS/MOS scores of sequences *i* and *j* are identical (their CIs overlap) whereas the evaluation in crowdsourced test says that they are different (their CIs do not overlap),

c) False Ranking, the worst error. It happens when the evaluation in laboratory test says that DMOS/MOS scores of the sequences *i* (*j*) are statistically better (according to their CIs) than *j* (*i*) whereas the evaluation in crowdsourced test says the opposite.

The error classification rates for DMOS/MOS quality, depth and comfort scores are presented in Table VIII. We used linear and cubic regression of scores from crowdsourced test, $P_{thresh}=Q_{thresh}=3$, with z-scores calculation, previously described case 2 (results for linear and cubic regression are same). In DMOS case, we had 84 and in MOS case, we had 88 video sequences.

TABLE VIII ERROR CLASSIFICATION FOR DMOS/MOS QUALITY, DEPTH AND COMFORT SCORES

	Quality (%)				Depth (%)				Comfort (%)			
	Correct	False Tie	False Differentiation	False Rank	Correct	False Tie	False Differentiation	False Rank	Correct	False Tie	False Differentiation	False Rank
DMOS	77.7682	4.9627	17.2691	0	72.2605	4.1308	23.6087	0	73.6948	6.6839	19.6213	0
MOS	82.1859	4.8480	12.9662	0	76.6208	5.0201	18.3592	0	74.8135	7.8887	17.2978	0

V. DISCUSSION

From the results in subsection IV.B it can be concluded that by using the proposed framework of the *Crowd3D* method it is possible to obtain similar DMOS/MOS quality scores as in laboratory experiments, provided all ARMs implemented and explained earlier are used. E.g. for case 2 - $P_{thresh}=Q_{thresh}=3$, with z-scores calculation, cubic fit, Pearson's and Spearman's correlations between crowdsourced and laboratory tests are about 0.94 for DMOS and 0.96 for MOS scores. However, correlation is somewhat lower for DMOS/MOS depths and comfort scores (for case 2 it is 0.89/0.93 and 0.89/0.91 for DMOS/MOS depth and comfort scores). Lower scores for comfort and depth can be due to the several reasons, which are very difficult to control in crowdsourced tests: different illumination conditions, different 3D monitor type, different monitor settings. Also, depth and comfort scores, as added grades in 3D subjective experiments, may require the use of different subjective assessment approaches (in our work we have used ACR-HR). Possibly, observers may be more uncertain when evaluating depth and comfort, than generic video quality as those two quality dimensions are harder to define.

When comparing DMOS/MOS scores between themselves, c.f. Table IV, it can be noticed that in the crowdsourced quality evaluations, inter-correlation between different DMOS scores is higher, comparing with laboratory results. This could be because

depth and comfort scores were not easily understandable to the observers as quality scores, which may have made quality, depth and comfort scores more similar. Possibly, this could then influence negatively the grading of the contents in the depth and comfort dimensions, resulting in unreliable DMOS scores for these two quality indicators. However, as was the case in the laboratory test, also for the crowdsourced data DMOS scores for quality and comfort have the highest correlation. In [35] authors have presented a comparison between "Visual quality", "Visual Discomfort" and "Sense of presence" gradings (and 2 viewing distances) obtained using the NAMAS1-COSPAD dataset [22]. For visual quality and sense of presence they used ACR scale, while for visual discomfort they used "Degradation Category Rating" scale [31]. They concluded that the different scales they used have high correlation: Pearson's correlation of 0.9 for visual quality - visual discomfort and 0.93 for visual quality - sense of presence grade pairs. This might show that general video quality scale is sufficient for evaluating side-by-side video experience, with the characteristics similar to that of NAMAS1-COSPAD dataset (mainly coding and spatial resolution reduction distortions). However, in our subjective experiment we had more different distortion types (some of which are specific for 3D distortion types), so those scales should represent more different grades.

When comparing error classification for DMOS/MOS quality, depth and comfort scores, Table VIII, again it can be seen that highest correct classification rate was obtained for DMOS/MOS quality scores (between laboratory and crowdsourced tests), lower in the case of DMOS/MOS depth and comfort scores.

Next we compare our results with some other similar performing 3D subjective evaluation tests. In [14] (between crowd-based and lab-based test), compared to our results on the agreement between laboratory and crowdsourced originated DMOS grades, that paper reports similar values for OR, only ANOVA test in [14] calculated 100% of correct estimation. Concerning error classification, we also obtained similar results for percentage of correct classification for quality scores (about 78% for DMOS and 82% for MOS). Spearman's correlation between crowd-based and lab-based evaluations in [14] was above 0.97. Similar conclusions can be drawn from [20] (3 different laboratories; similar setup for tests as in [14]), only OR results from [20] were worse than in our comparison. ANOVA test in this case estimated similar results (88.89% - 98.61% of correct estimation, depending on laboratory). Spearman's correlation between laboratories in [20] was 0.9340-0.9399. In [21] (3 different laboratories; authors tested 10 degradation types from NAMAS1-COSPAD dataset) Spearman's correlation between laboratories was 0.9634-0.9811.

When comparing number of grades per sequence with [8], it should be noted that 2D image quality assessment can be done more easily than 3D video quality assessment, especially for crowdsourced tests.

When comparing TV sets only and inter-correlation between DMOS/MOS grades, Table IV, it can be seen that better differentiation between quality, depth and comfort scores were obtained than crowdsourced test with $P_{thresh}=Q_{thresh}=3$. Also, from Table V best correlation for Pearson's and Spearman's quality DMOS, best correlation for Spearman's comfort DMOS, nearly the best for Pearson's comfort DMOS (slightly better is case 7. - 20-35 y–Cubic fit), lowest RMSE for quality DMOS and nearly the lowest for comfort DMOS (slightly better is case 7. - 20-35 y–Cubic fit) were also obtained using only grades from TV sets (better than all other tested groups). Correlation for depth DMOS grades and RMSE were better in overall results (case 1, with $P_{thresh}=Q_{thresh}=2$, with z-scores calculation). For MOS scores, Table VI, TV sets have highest correlation and lowest RMSE for comfort scores, second highest correlation and second lowest RMSE for depth scores (first is case 1, $P_{thresh}=Q_{thresh}=2$, with z-scores calculation). For MOS quality scores, tested case 4 ($P_{thresh}=Q_{thresh}=4$, with z-scores calculation) has the best correlation and lowest RMSE. This may lead to the conclusion that in crowdsourced evaluations, it is better to use TV sets only; possibly, general 3D quality and comfort grades on monitors and laptops are more diverse than on TV sets only, comparing with laboratory evaluation (and depth and comfort mixed with quality grades; comparing Table IV). Another reason may be because laboratory evaluation was also made only on TV set. Similar conclusion can be seen in in ITU P.914 [3], where comparison between different TV sets usually has higher correlation than between TV and laptop.

Other factors, like age, gender and active glasses devices, did not have important influence, when comparing with overall results. Although somewhat better differentiation between quality, depth and comfort scores were obtained (Table IV), still those subset of results have somewhat lower or similar correlation with laboratory tests, when comparing with overall results for both DMOS/MOS scores (e.g. with case 2). Lower correlation in some cases can be also due to the lower number of observers, when testing those specific factors (Table V and Table VI, min and max number of observers).

When comparing overall results, with and without z-scores calculation, it can be seen that higher correlation with laboratory test is obtained using z-scores. However, from results without z-scores calculation it can be seen that most observers were removed due being outliers from mean for depth grades (12 or 15 respectively for DMOS/MOS), then for comfort grades (11 or 12 respectively for DMOS/MOS), and lowest number for quality grades (7 or 8 respectively for DMOS/MOS). This may be due to the observers having highest uncertainty in giving depth grades.

From Table VIII it can be seen that false differentiation error is much higher than the false tie error for video quality, depth quality, and visual comfort DMOS/MOS grades. From definition, this means that there exist more cases where laboratory scores are identical, whereas in crowdsourced test those cases are different. This can be explained due to the larger CI in laboratory test because in laboratory test there were on an average 18 grades per video sequence, while in crowdsourced test on average, each degraded video sequence was graded about 34.8 times.

When comparing IP addresses (that application monitored and saved) with countries that observers told they live, generally answers are correct. For 6 observers, wrong IP addresses could be due to proxy servers; removing those observers did not improve Pearson's and Spearman's correlations.

When comparing cases 9 and 10 (ARMs that have been used are removed: false observers in case 9 and false and double in case 10), correlation is similar in case 9 or somewhat smaller comparing to e.g. case 2 probably because only 12 sessions were added. Correlation did not significantly drop even in case 10 (false and double evaluations – giving 220 evaluations, comparing with 139 valid). This may be explained because initially, we used prescreening of all observers (phase 1), which probably removed many false observers anyway (and those scores we cannot compare as we used 5 video sequences in phase 1 – not the same as in phase 2). In the case all 283 evaluations are considered (only MOS scores can be calculated in this case because grades from original video sequences are generally not reported), correlation results are similar like case 10, because usually only a few grades (or even none) have been acquired from those unfinished evaluations. Another reason might also be, as stated earlier, lower number of 3D equipment among general population, which makes cheaters probably more reluctant to participate in the study.

VI. Conclusion

This paper proposes a new method for crowdsourced subjective 3D video quality assessment – *Crowd3D*. A comparison with the results obtained in controlled laboratory-based studies is also given.

It can be concluded that by using the proposed framework of the *Crowd3D* method it is possible to obtain grades with high correlation with laboratory collected grades for quality scores, but somewhat lower correlation for depth and comfort scores. Reasons for that could be different: possibly too low number of the observers (especially for depth and comfort scores), depth and comfort scores not easily understandable to the observers as quality scores, and finally test equipment and test conditions which may have a stronger effect on depth and comfort grades, than on quality grades. Although the proposed crowdsource application uses several mechanisms to check and improve the reliability of the results, the influence of external factors such as monitor type, illumination quality and its colour temperature, cannot be removed entirely. Further research may be needed to fully understand the new quality dimensions associated with 3D video and respective scores (depth, comfort), by using similar equipment in different conditions in both laboratory and crowdsourced environments, using more observers and maybe changing the methodology to be used in 3D video subjective tests (double stimulus instead of single stimulus, maybe even using different description of those additional scores, etc.). Future 3D crowdsourced evaluations could also include approximate information about distance from the screen, as it could give information, together with screen size, about whether is screen or some of its part outside the zone of visual comfort, and its influence on subjective grades.

As an additional contribution to this research area, the video sequences used in this work and related DMOS/MOS scores for quality, depth and comfort (using case 2, *DMOS/MOS* scores, $P_{thresh}=Q_{thresh}=3$ with z-scores calculation) are made publicly available. The whole dataset can be found at repository [36] and includes the compressed video sequences together with the collected grades information and Crowd3D application source code.

Acknowledgement

Authors would like to thank the European COST Action IC1105, 3DConTourNet for the active support and cooperation.

References

[1] ITU-R BT.500-13 "Methodology for the subjective assessment of the quality of television pictures", International Telecommunication Union/ITU Radiocommunication Sector, 2012.

[2] ITU-R BT.2021 "Subjective methods for the assessment of stereoscopic 3DTV systems", International Telecommunication Union/ITU Radiocommunication Sector, 2012.

[3] ITU-R P.914 "Display requirements for 3D video quality assessment", 2016

[4] ITU-R P.915 "Subjective assessment methods for 3D video quality", 2016

[5] ITU-R P.916 "Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video", 2016

[6] M Loncaric et al., "Testing picture quality in HDTV systems", ELMAR, 2008 50th International Symposium , September 2008, pp. 5-8.

[7] E. Dumic, S. Grgic, K. Sakic, D. Frank, "Subjective Quality Assessment of H.265 versus H.264 Video Coding for High-Definition Video Systems", 13th International Conference on Telecommunications ConTEL 2015, ConTEL 2015 Proceedings, July 2015, pp. 1-7.

[8] D. Ghadiyaram, A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality", IEEE Trans. on Image Processing, Vol. 25, No. 1, November 2015, pp. 372-387

[9] D. Ghadiyaram, A. C. Bovik, "Crowdsourced study of subjective image quality", 2014 48th Asilomar Conference on Signals, Systems and Computers, November 2014, pp. 1-5

[10] K. Sakic, E. Dumic, S. Grgic, "Crowdsourced Subjective Video Quality Assessment", The 21st International Conference on Systems, Signals and Image Processing – IWSSIP, IWSSIP 2014 Proceedings, May 2014, pp. 223-226.

[11] Microworkers, March 2016. [Online]. Available: http://microworkers.com

[12] Amazon Mechanical Turk, March 2016. [Online]. Available: http://mturk.com

[13] E. Dumic, S. Grgic, K. Sakic, P. M. R. Rocha, L. A. da Silva Cruz "3D video subjective quality: a new database and grade comparison study", Multimedia tools and applications (1380-7501) (2016), January 2016, pp. 1-23.

[14] P. Hanhart, P. Korshunov, T. Ebrahimi "Crowd-based quality assessment of multiview video plus depth coding", IEEE International Conference on Image Processing, Paris, France, October 2014, pp. 743-747

[15] T. Hoßfeld et al., "Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing", IEEE Transactions on Multimedia, Vol. 16, No. 2, February 2014, pp. 541-557

[16] T. Hoßfeld, J. Redi, "Journey through the crowd: Best practices and recommendations for crowdsourced QoE", Quality of Multimedia Experience (QoMEX) 2015, May 2015, pp. 1-2

[17] R. G. Kaptein, A. Kuijsters, M. T. M. Lambooij, W. A. IJsselsteijn, I. Heynderickx, "Performance evaluation of 3D-TV systems", Proceedings of SPIE Image Quality and System Performance V, SPIE Vol. 6808, 680819, 2008

[18] M. Lambooij, W. IJsselsteijn, D. G. Bouwhuis, I. Heynderickx, "Evaluation of Stereoscopic Images: Beyond 2D Quality", IEEE Transactions on Broadcasting, Vol. 57, No. 2, June 2011, pp. 432–444.

[19] M. T. M. Lambooij, W. A. IJsselsteijn, I. Heynderickx, "Visual Discomfort in Stereoscopic Displays: A Review", Stereoscopic Displays and Virtual Reality Systems XIV, SPIE Vol. 6490, 64900I, 2007

[20] P. Hanhart, N. Ramzan, V. Baroncini, T. Ebrahimi, "Cross-lab Subjective Evaluation of the MVC+D and 3D-AVC 3D Video Coding Standards", 6th International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, September 2014, pp. 183-188

[21] M. Barkowsky, J. Li, T. Han, S. Youn, J. Ok, et al., "Towards standardized 3DTV QoE assessment: Cross-lab study on display technology and viewing environment parameters", SPIE Electronic Imaging, , San Franscisco, United States. 8648, February 2013, pp.864809-864809

[22] M. Urvoy et al., "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences", Quality of Multimedia Experience (QoMEX 2012), Melbourne, Australia, July 2012, pp. 109-114

[23] M. Lambooij, M. Murdoch , W.A. IJsselsteijn, I. Heynderickx, "The impact of video characteristics and subtitles on visual comfort of 3D TV", Displays, Vol. 34, No. 1, January 2013, pp. 8–16

[24] J. Paulus, G. Michelson, M. Barkowsky, J. Hornegger, B. Eskofier, M. Schmidt, "Measurement of Individual Changes in the Performance of Human Stereoscopic Vision for Disparities at the Limits of the Zone of Comfortable Viewing", 2013 International Conference on 3D Vision, 2013, pp. 310–317

[25] P. Reichl, S. Egger, S. Möller, K. Kilkki, M. Fiedler, T. Hossfeld, C. Tsiaras, A. Asrese, "Towards a comprehensive framework for QOE and user behavior modelling", 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), May 2015, pp. 1-6

[26] Google Chrome based application for laboratory subjective assessments, March 2016. [Online]. Available: crowd3d.co.it.pt/suis3d

[27] Mozilla Firefox based application for laboratory subjective assessments, March 2016. [Online]. Available: crowd3d.co.it.pt/suis3d_webm

[28] Application for Phase 1 of the crowdsourced subjective 3D video quality assessment, March 2016. [Online]. Available: http://crowd3d.co.it.pt/test/suis.php

[29] Application for Phase 2 of the crowdsourced subjective 3D video quality assessment, March 2016. [Online]. Available: http://crowd3d.co.it.pt/crowd3d/suis.php

[30] Dataset repository, June 2016. [Online]. Available: ftp://ftp.ivc.polytech.univ-nantes.fr/NAMA3DS1_COSPAD1/Avi_videos/HRC_00_Reference/

[31] ITU-T, "Recommendation P.910, Subjective video quality assessment methods for multimedia applications", 2008.

[32] H.R. Sheikh, "Image Quality Assessment Using Natural Scene Statistics," Ph.D. dissertation, University of Texas at Austin, (May 2004)

[33] ITU-T J.149, "Method for specifying accuracy and crosscalibration of Video Quality Metrics (VQM)," International Telecommunication Union, March 2004.

[34] P. Hanhart and T. Ebrahimi, "On the evaluation of 3D codecs on multiview autostereoscopic display", 4th IEEE International Workshop on Hot Topics in 3D (Hot3D), July 2013., pp. 1-2

[35] K. Brunnström, I. V. Ananth, C. Hedberg, K. Wang, B. Andrén and M. Barkowsky, "36.4: Comparison between Different Rating Scales for 3D TV. SID Symposium Digest of Technical Papers", Vol. 44, No. 1, June 2013, pp. 509–512.

[36] Dataset repository, June 2016. [Online]. Available: http://beam.to/datasets

Emil Dumic received his MSc and PhD degree from University of Zagreb, Faculty of Electrical Engineering and Computing in 2007 and 2011 respectively.
He is an Assistant Professor at the University North, Department of Electrical Engineering. 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.
Kresimir Sakic received his M. E.E. and PhD degree from University of Zagreb, Faculty of Electrical Engineering and Computing in 2006 and 2016 respectively.
He works as a Senior Broadcasting Planning Expert in the Croatian Regulatory Authority for Network Industries in Zagreb. His current research interests include subjective and objective assessments of 2D and 3D video databases, video coding performance assessment, radio-wave propagation, transmitter coverage predictions and comparison with field measurements, interference analysis, etc.
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 Instituto de Telecomunicações in 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.