Impact of interactivity on the assessment of quality of experience for light field content

Irene Viola, Martin Řeřábek and Touradj Ebrahimi
Multimedia Signal Processing Group
EPFL, Lausanne, Switzerland

***Abstract*—The recent advances in light field imaging are changing the way in which visual content is captured, processed and consumed. Storage and delivery systems for light field images rely on efficient compression algorithms. Such algorithms must additionally take into account the feature-rich rendering for light field content. Therefore, a proper evaluation of visual quality is essential to design and improve coding solutions for light field content. Consequently, the design of subjective tests should also reflect the light field rendering process. This paper aims at presenting and comparing two methodologies to assess the quality of experience in light field imaging. The first methodology uses an interactive approach, allowing subjects to engage with the light field content when assessing it. The second, on the other hand, is completely passive to ensure all the subjects will have the same experience. Advantages and drawbacks of each approach are compared by relying on statistical analysis of results and conclusions are drawn. The obtained results provide useful insights for future design of evaluation techniques for light field content.**

Keywords*—*light field, subjective evaluation, interactivity, image coding, image compression.

I. Introduction

New acquisition technologies to capture, process and visualize Light Field (LF) contents are shaping the future of photography towards new feature-rich representations. The idea behind LF imaging is to provide more complete representation of a scene by recording the direction of light, along with its intensity. This can be achieved with different acquisition technologies, such as, among others, multi-camera arrays or hand-held devices. However, the enhanced features of LF imaging come with a substantial increase in the volume of data generated in the acquisition process. More specifically, the availability of LF cameras in the consumers' market presents new challenges in terms of storage, representation and visualization of the acquired data.

Designing new solutions, to face the challenges LF imaging poses, cannot forgo the importance of evaluating them in a reliable and reproducible way. In particular, subjective assessment of visual quality is of paramount importance to evaluate the impact of compression, representation, and rendering models on user experience. Various examples of subjective quality evaluation methodologies can be found in literature. Paudyal et al. analyse the impact of watermarking on visual quality of LFs using Absolute Category Rating (ACR) [1]. Their effort focuses on the relationship between watermark strength and visual quality. In their previous work, the authors evaluate compression solution through objective and subjective quality assessment, in the framework of ICME 2016 Grand Challenge [2]. The evaluation is performed on several viewpoints extracted from LF contents and displayed as still images alongside with their uncompressed references, using a methodology based on Double Stimulus Continuous Quality Scale (DSCQS). Darukumalli et al. and Kara et al. investigate the quality of experience associated with LF displays, and its relationship with angular resolution and zooming levels [3], [4]. To do so, they use ACR and Degradation Category Rating (DCR).

QoMEX2017 – Erfurt, Germany; 978-1-5386-4024-1/17/$31.00 ©2017 IEEE

Design of subjective quality evaluation is a delicate task that requires careful considerations, since it has an influence on the statistical relevance of the results. Moreover, a well designed quality evaluation experiment should take into account how the end user engages with the content. This is especially important when evaluating LF contents. Thus, rendering techniques for LF contents and subjective evaluation of quality cannot be considered as independent problems.

LF contents offer a wide range of possibilities for rendering. For example, different viewpoints can be accessed, digital refocusing can be applied, super-resolution algorithms can be used to increase the resolution of the image, and so on. The most natural way for the user to exploit these possibilities is by interacting with the content. Indeed, being able to change the appearance of the scene that has been acquired is a desirable feature, one that is already implemented in widespread applications such as Instagram or Facebook. From this perspective, interactive methodologies for subjective assessment should be actively deployed since they give a more accurate depiction of how the user consumes and engages with the content. One example of interactive methodology is proposed by the authors in [5], where users can access different viewpoints and change the focal point by interacting with the content through a GUI. However, one significant shortcoming of the interactive approach is the lack of control on what users are visualising and thus what is being rated. Since each subject decides autonomously which viewpoint to display and for how long, there is little control over the number of viewpoints that each subject is examining, nor there is guarantee that the viewpoints selected by different subjects are the same.

An alternative way to evaluate visual quality of LF contents would be to use a passive approach, where the subjects are presented with a pre-recorded animation displaying different viewpoints. Such an approach guarantees that each subject sees the identical set of viewpoints under the same conditions. However, to yield reliable results, a number of parameters should be carefully selected, such as the optimal framerate and the number of viewpoints to be presented to the subject. Moreover, a passive approach disregards the interactive nature

(a) Bikes (b) Stone_Pillars_Outside (c) Fountain_&_Vincent_2 (d) Friends_1

Fig. 1: Central viewpoint image from each content used in our experiment.

TABLE I: Values of refocusing slopes for each content.

	Slopes										
Content	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Bikes	-10	-8	-6	-4	-2	0	2	4	6	8	10
Stone_Pillars_Outside	-10	-8	-6	-4	-2	0	2	4	6	8	10
Fountain_&_Vincent_2	-10	-8	-6	-4	-2	0	2	4	6	8	10
Friends_1	-5	-4	-3	-2	-1	0	1	2	3	4	5

of LF contents, and thus does not always faithfully represent the average user experience in consuming the LF content.

In this paper, we compare results of subjective assessments of visual quality obtained by using two methodologies, one that enforces interaction with the content, and one that favors an automated presentation. For the first methodology, a controlled lab environment was adopted, while for the second methodology, due to time and costs constraints, a crowdsourcing tool was deployed.

The remainder of the paper is organized as follows. Details on how the experiments were designed and carried out are presented in section II. Then, statistical metrics and analysis tools are introduced in section III. Finally, results from the comparison are discussed in section IV, and conclusions are drawn in section V.

II. Experimental test design

This section describes how the subjective evaluations were designed. More specifically, the creation of the stimuli for both tests is outlined. A description of the interactive subjective methodology, along with the testing environment, is presented. Then, the passive subjective methodology is described in details. A summary of the specifications for the two methodologies can be found in Table II.

A. Data preparation

Four LF images, acquired by a Lytro Illum camera, were selected from a publicly available LF image dataset [6]. In particular, contents *Bikes*, *Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* were selected for the experiments. Thumbnails for each content are depicted in Figure 1. Following ITU Recommendations [7], the images were carefully selected in order to provide a wide range of scenarios, including details that would prove critical for the compression algorithms.

The lenslet images were processed using the LF MATLAB toolbox [8], [9] to obtain the collection of viewpoints needed for the subjective tests. Additionally, eleven refocused images were created for each content, using a modified version of the toolbox function *LFFiltShiftSum*. For our tests, it was decided to sum images from index 3 to index 13 (11×11 images) to have a larger depth of field than that obtained by shifting and summing all of the viewpoints. The values of the slopes used to shift the viewpoints are summarized in Table I. The slopes were selected to assure gradual transition between refocusing on the foreground and on the background with respect to semantically relevant objects in each content.

The uncompressed reference was obtained by preprocessing the raw sensor data through devignetting, demosaicing, clipping to 8 bits, transforming to a collection of viewpoints and applying color and gamma corrections. The reference was obtained from the lenslet image in RGB 444, without any chroma subsampling. This reference was selected to have a proper comparison with acquisition data obtained with minimal pre-processing. For this reason, chroma subsampling was not applied on the reference, since it alters the data.

Five compression algorithms were used to create the data to evaluate the two methodologies. Three anchors were created by the authors using HEVC encoding (x265 implementation), whereas two others were taken from literature [10], [11]. Each compression scheme was given a label for easier identification. A summary of the compression schemes can be found in Table III. The compression algorithms were evaluated on four bitrates (corresponding to four compression ratios), namely $R1 = 1$ bpp $(10:1)$, $R2 = 0.5$ bpp $(20:1)$, $R3 = 0.25$ bpp $(40:1)$, $R4 = 0.1$ bpp $(100:1)$. The compression ratios were computed as ratios between the size of the uncompressed raw images in 10bit precision and the size of the compressed bitstreams.

B. Interactive methodology

To perform the interactive visual assessment, a recently introduced methodology for evaluation of plenoptic content was selected [5]. The methodology is based on Double Stimulus Impairment Scale (DSIS) [7].

Participants were asked to interact with the LF images and rate the level of impairments of the test LF image with

TABLE II: Test environments and specifications.

Approach	Environment	No. subjects	Methodology	No. viewpoints	No. refocused views	fps	Median age
Interactive	Controlled lab setting	24	DSIS	169	11	-	25
Passive	Semi-controlled crowdsourcing	24	DSIS	97	11	30	22

TABLE III: Summary of compression schemes.

Proponents	Description
P01	Lenslet image compressed using HEVC intra (software x265).
P02	Lenslet image compressed using HEVC intra with LLE and SS (software HM-14.0) [10].
P03	Lenslet image compressed using intermediate transformation to viewpoints and HEVC (software JEM 2.0) [11].
P04	Chroma subsampling of the lenslet image and compression of viewpoints through pseudo-temporal sequence using HEVC (software x265).
P05	Compression of viewpoints through pseudo-temporal sequence using HEVC (software x265).

respect to the reference, on a scale from 1 (*Very annoying*) to 5 (*Imperceptible*). Each LF image was presented together with the uncompressed reference in a side-by-side fashion. The position of the reference was set to either left or right for each experiment, and participants were informed about its location on the screen. For each stimulus, the central viewpoint image from the LF image was displayed. By clicking inside the displayed image and dragging the mouse, the other viewpoints from the LF image were accessed and displayed. Each image was displayed in its native resolution of 625×434 pixels. A total of 13×13 viewpoints were accessible. The refocused images were accessible through a slider shown at the bottom of each stimulus.

To avoid the involuntary influence of external factors and to ensure the reproducibility of results, the laboratory for subjective video quality assessment was set up according to ITU-R Recommendation BT.500-13 [7]. Professional Eizo ColorEdge CG301W 30-inch monitors with native resolution of 2560×1600 pixels were used for the tests. The monitors were calibrated using an i1Display Pro color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level of 0.2 cd/m^2. The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. The illumination level measured on the screens was 15 lux. The distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content, conforming to requirements in ITU-R Recommendation BT.2022 [12].

Before the experiments, a training session was organized to allow participants to get familiar with artefacts and distorsions in the test images. Five training samples were manually selected by expert viewers. The training samples were created by compressing other content on various bitrates. The content used for the training was selected from the same LF image database used for the test images [6]. The training samples were presented along with the uncompressed reference, exactly as they were shown in the tests.

The experiment was split into two sessions. In each session, 40 stimuli were shown side by side with the uncompressed reference, corresponding to approximately 20 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all the sessions, thus evaluating the entire set of stimuli. A break of ten minutes was enforced between the sessions to avoid fatigue. Before the test, one dummy sample was inserted to ease the participants into the task. The resulting scores from dummy stimuli were not included in the results.

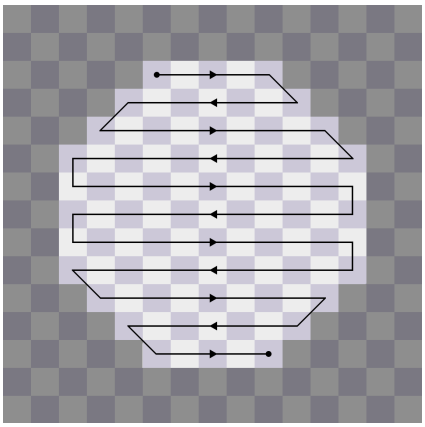


Fig. 2: Ordering of the views for animation for passive methodology.

A total of 24 subjects (19 males and 5 females) participated in the experiment, for a total of 24 scores per stimulus. Subjects were between 18 and 35 years old, with an average of 24.79 and a median of 25 years of age. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts.

C. Passive methodology

The passive visual assessment of quality was carried on using a methodology based on DSIS [7]. To perform the tests, the QualityCrowd 2 framework [13] was used. However, it should be noted that all the participants performed the tests in the same environment at the same time, with equal lighting conditions, using the same display model and the same screen resolution. The participants were shown the LF content as a video sequence navigating between the viewpoints and the refocused images. Each stimulus was displayed alongside with the uncompressed reference, in a side by side fashion. The subjects knew in advance on which side of the screen the reference was displayed.

Due to distortions caused by the lenslet structure, several viewpoints presented artefacts independent from the coding procedure, and thus had to be discarded. Only a subset of 97 out of 225 viewpoints was chosen to be displayed, in order not

TABLE IV: Selected settings for AVC coder for passive methodology.

-r 30 -s <*size*> -f rawvideo -pix_fmt yuv420p -i <*input*> -c:v libx264 -profile:v high -x264opts no-scenecut:no-deblock:pass=1 -b:v 8M tmp.mp4
-r 30 -s <*size*> -f rawvideo -pix_fmt yuv420p -i <*input*> -c:v libx264 -profile:v high -x264opts no-scenecut:no-deblock:pass=2 -b:v 8M <*output*>

to affect the rating. Ten viewpoints per second were displayed, to ensure a smooth transition of the different viewpoints. The viewpoints were accessed from top to bottom and from left to right and right to left in alternate order (see Figure 2). At the end of the viewpoint animation, the eleven refocused images were displayed with a framerate of four refocused images per second, going from foreground to background and from background to foreground. The animation setup was chosen and validated by expert viewers in order to mimic the parallax effect, as well as to mimic the refocusing effect that occurs when trying to change the focal point. The total length of the animation for each stimulus was 14 seconds. Since there is no browser video plugin capable of reliable real-time decoding and displaying for HEVC, the animations were encoded with AVC. A two-pass encoding was used and the deblocking filter was disabled to ensure transparency and to preserve the original blockiness artefacts when encoded at low bit rates. Expert viewing session conducted prior to the main subjective assessment concluded that the AVC video encoding was visually lossless, and thus would not influence in any way the final scoring. Selected settings for AVC coder are summarised in Table IV.

Test subjects were asked to rate the level of impairment of the test stimuli when compared to the uncompressed references. The rating was performed on a scale from 1 (Very annoying) to 5 (Imperceptible). Before the experiment, a training session was organized to allow participants to get familiar with artefacts and distorsions in the test images. Five training samples were manually selected by expert viewers. To help subjects localize and identify compression artefacts in the fast-paced video, the same content used in the test was selected for the training. The training samples were presented along with the uncompressed reference, exactly as they were shown in the test.

The experiment was split into two sessions. In each session, 40 stimuli were shown side by side with the uncompressed reference, corresponding to approximately 20 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all the sessions, thus evaluating the entire set of stimuli. A break of ten minutes was enforced between the sessions to avoid fatigue.

A total of 24 subjects (22 males and 2 females) participated in the experiment, for a total of 24 scores per stimulus. Subjects were between 18 and 35 years old, with an average of 22.79 and a median of 22 years of age.

III. Statistical analysis

Outlier detection and removal was performed on the results, independently for each methodology, according to the ITU Recommendations [7]. One outlier was detected in results obtained using the interactive methodology, whereas no outlier was found in the results from the passive methodology. This led to 23 scores per stimulus for the first method, and 24 scores per stimulus for the second. After outlier removal, the mean opinion score (MOS) was computed for each stimulus, independently for each methodology. The corresponding 95% confidence intervals (CIs) were computed assuming a Student's t-distribution.

Following the ITU Recommendations [14], several fittings were applied to the MOS values from the two different methodologies. In particular, first order and third order fittings were used to compare the MOS values. Absolute prediction error (RMSE), Pearson Correlation Coefficient (PCC), Spearman's Rank Correlation Coefficient (SRCC) and Outlier Ratio (OR) were computed for accuracy, linearity, monotonicity and consistency, respectively.

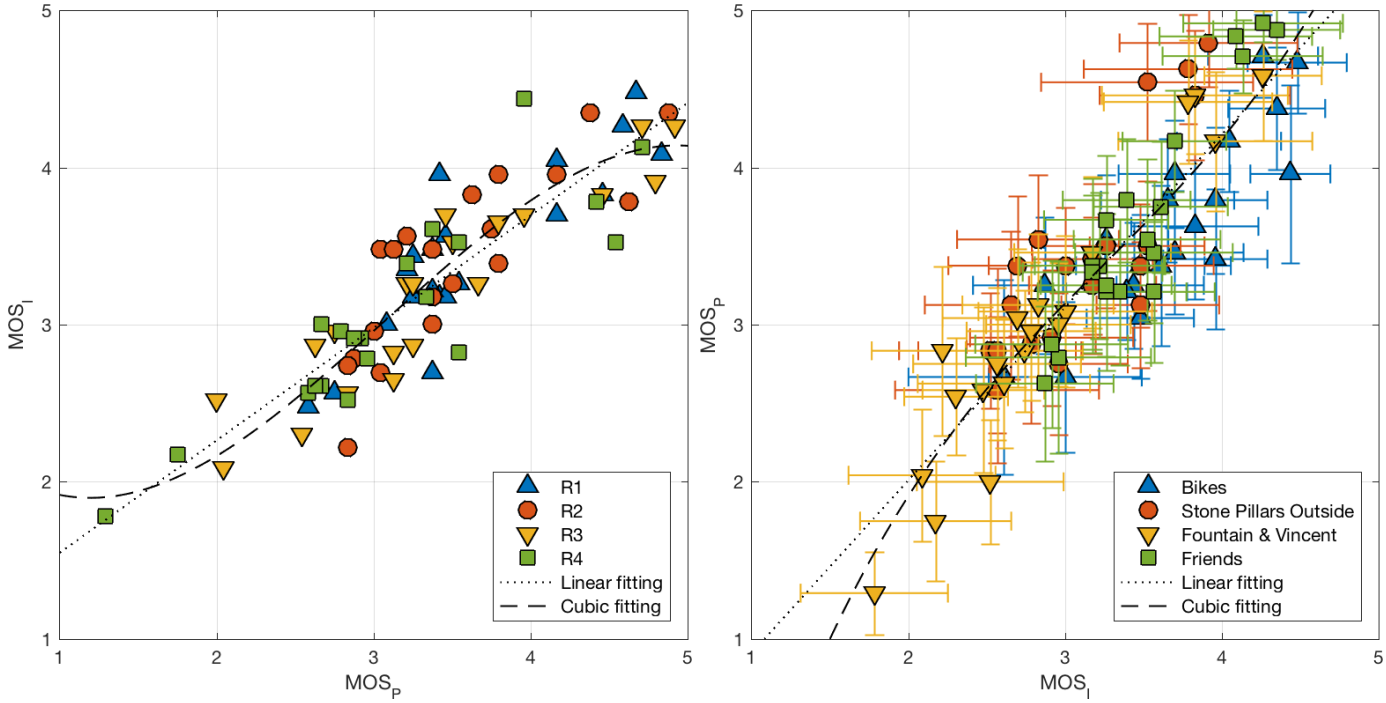
A multiple comparison test was performed at a 5% significance level on the raw scores, to determine, for each stimulus, whether the MOS values obtained with the two methodologies were significantly different, and the percentage of correct estimation, underestimation and overestimation were computed. Additionally, the classification errors were computed using the same multiple comparison test to see if the results obtained with the two methodologies lead, for each pair of stimuli, to the same conclusions [15]. In this case, three types of error can be distinguished: false ranking, false differentiation and false tie. False ranking is the most offensive error, and occurs when the first methodology says that situation i is better than situation j, whereas the second methodology says the opposite. False differentiation occurs when the first methodology says that situation i and j are different, whereas the second methodology says they are the same. False tie occurs when the first methodology says two situations are the same, whereas the second methodology says they are different.

Finally, one-way and multi-way ANOVA tests were performed to assess the influence of the methodology on the results, and in particular whether the two methodologies lead to significantly different results.

IV. Results and discussion

Figure 3 shows the scatter plots comparing the MOS values obtained with the two tested methodologies. On the right, the horizontal and vertical bars represent the CIs corresponding to results obtained with interactive and passive methodologies, here denominated I and P, respectively. To improve visualization, the points are colored based on compression ratio or content. Linear and cubic regressions are shown for both comparisons. Table V shows the performance indexes computed on the data. The indexes are computed on the data pairs $[\widehat{MOS}_A, MOS_B]$ where $A, B = I, P$. $\widehat{MOS}_A$ are the MOS scores obtained with methodology A with no fitting, linear fitting and cubic fitting, and MOS_B are the MOS scores obtained with methodology B.

Ideally, a 45° line would indicate that the two methodologies give the same MOS values for the same condition. However, as it is visible in Figure 3, the points are not aligned along the $y = x$ line. In particular, linear regression performed on MOS_P has a slope of 0.716 and an intercept



(a) MOS_I as function of MOS_P.

(b) MOS_P as function of MOS_I.

Fig. 3: Comparison of MOS values obtained with the different methodologies, along with linear and cubic fittings. Points are differentiated by compression ratio (a) and by content (b).

TABLE V: Performance indexes.

	$[\widehat{MOS}_P, MOS_I]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Correct Decision**	**False Ranking**	**False Differentiation**	**False Tie**
No fitting	0.8878	0.8876	0.3791	3.75%	100%	84.56%	0.00%	13.04%	2.41%
Linear fitting	0.8878	0.8876	0.2797	0.00%	100%	89.37%	0.00%	3.26%	7.37%
Cubic fitting	0.8957	0.8876	0.2708	0.00%	100%	88.80%	0.00%	0.82%	10.38%
	$[\widehat{MOS}_I, MOS_P]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Estimation**	**Correct Decision**	**False Ranking**	**False Differentiation**	**False Tie**
No fitting	0.8878	0.8876	0.3791	3.75%	100%	84.56%	0.00%	2.41%	13.04%
Linear fitting	0.8878	0.8876	0.3468	0.00%	100%	86.84%	0.00%	3.26%	9.91%
Cubic fitting	0.8895	0.8876	0.3444	0.00%	100%	89.97%	0.00%	6.42%	3.61%

of 0.832, which indicates that, on average, for the same stimulus subjects gave a higher rating when presented with passive methodology as opposed to interactive methodology. This is confirmed by the results of boxplot analysis on the two methodologies, which shows that, on average, results obtained with the passive methodologies tend to have higher ratings. This tendency can be explained considering that viewers are presented with a carefully selected subset of viewpoints in the passive experiments, which are less prone to lenslet-based artefacts, as opposed to the wider number of viewpoints viewers can access in the interactive experiments.

Cubic regression has a sigmoid shape in both $\widehat{MOS}_P$ and $\widehat{MOS}_I$, as confirmed by values obtained performing PCC and SRCC, which indicate a strong but not perfect linear correlation. Low values of RMSE and OR confirm the correlation between the two methodologies. Furthermore, there is no over- or under-estimation, as proven by correct estimation being 100%, which indicates that, for the same stimulus, there is no statistically significant difference between the scores obtained with one or the other methodology.

One-way ANOVA performed on stimuli grouped only by methodology shows that results obtained with the two methodologies, although highly correlated, are statistically significantly different ($p = 0.0005$). To further investigate the influence of the coding parameters on the scores, we performed multi-way ANOVA on the results, separately for different compression ratios, contents and codecs, respectively. Results show that, for compression ratios $R2$ and $R4$, the two methodologies are statistically equivalent at 5% significance level, whereas for the remaining compression ratios they are statistically significantly different ($p = 0.008$ and $p = 0.0046$ for $R1$ and $R3$, respectively). For content *Bikes*, the two methodology are statistically equivalent, whereas for the remaining contents the two methodologies are statistically different ($p = 0$, $p = 0.044$ and $p = 0.0131$ for contents *Stone_Pillars_Outside, Fountain_&_Vincent_2* and *Friends_1*,

respectively). Finally, multi-way ANOVA analysis on different codecs shows that $P5$ is the only codec for which the two methodologies provide statistically different results ($p = 0$).

The classification errors show that there is no false ranking, the most offensive error. However, results from false differentiation performed on $[\widehat{MOS}_P, MOS_I]$ with no fitting show that, on 13.04% of cases, passive methodology considers two stimuli as being statistically significantly different, whereas the interactive methodology does not differentiate them. The percentage thus shows that the passive methodology has more discriminating power when compared to the interactive methodology. This is confirmed by comparing the CIs obtained with the two methodologies: on average, CIs obtained with passive methodology are 8.66% smaller. In other words, the standard error obtained with interactive methodology on 23 subjects would be equivalent to the standard error obtained with passive methodology on 20.13 subjects. Conversely, when using the interactive methodology, 27.42 subjects would be needed to obtain the same standard error provided by the passive methodology on 24 subjects.

It should be noted that, whereas the interactive evaluation has been conducted in a lab setting compliant with the guidelines set by ITU Recommendations [7], the passive evaluation has been carried out using crowdsourcing, which is usually associated with less reliable scores. However, several studies have proven the efficacy of crowdsourcing-based tests [16], [17]. Moreover, while crowdsourcing is usually linked to larger standard errors, due to variability of conditions, the opposite has been observed in our experiment. It shows that the passive approach contributed to lower the variance of the scores, in spite of the impact crowdsourcing might have in increasing the variance of the results.

V. Conclusions

In this paper we described the results of comparing two different approaches for subjective evaluation of visual quality for light field images. The statistical analysis performed on the results showed that the two approaches are highly correlated, although not statistically equivalent, and lead to similar ratings. However, we found that the interactive approach leads to larger confidence intervals in the corresponding scores, due to lack of control over the number of viewpoints that each participant visualises. Conversely, the passive approach, although conducted in a less controlled environment, showed a significant reduction in confidence intervals, and thus an increased discriminative power.

Interaction is a very desirable feature in light field quality assessment. Future design of evaluation methodologies for light field content should consider improving consistency for interactive testings, for example by merging the two approaches, or by adding tracking of user interaction to analyse patterns in user behaviour.

Acknowledgment

This work has been conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021_159575) project Light field Image and Video coding and Evaluation (LIVE) and also in the framework of ImmersiaTV under the European Unions Horizon 2020 research and innovation programme (grant agreement no. 688619) funded by Swiss State Secretariat for Education, Research and Innovation SERI.

References

[1] P. Paudyal, F. Battisti, A. Neri, and M. Carli, "A study of the impact of light fields watermarking on the perceived quality of the refocused data," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*. IEEE, 2015, pp. 1–4.

[2] I. Viola, M. Rerabek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and subjective evaluation of light field image compression algorithms," in *32nd Picture Coding Symposium (PCS)*, 2016.

[3] S. Darukumalli, P. A. Kara, A. Barsi, M. G. Martini, and T. Balogh, "Subjective quality assessment of zooming levels and image reconstructions based on region of interest for light field displays," in *2016 International Conference on 3D Imaging (IC3D)*, 2016.

[4] P. A. Kara, M. G. Martini, P. Kovacs, S. Imre, A. Barsi, K. Lackner, T. Balogh *et al.*, "Perceived quality of angular resolution for light field displays and the validity of subjective assessment," in *2016 International Conference on 3D Imaging (IC3D)*, 2016.

[5] I. Viola, M. Řeřábek, and T. Ebrahimi, "A new approach to subjectively assess quality of plenoptic content," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2016, pp. 99 710X–99 710X.

[6] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

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

[8] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun 2013.

[9] ——, "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, Feb. 2015.

[10] R. Monteiro, L. Lucas, C. Conti, P. Nunes, N. Rodrigues, S. Faria, C. Pagliari, E. Silva, and L. D. Soares, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2016, pp. 1–4.

[11] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–4.

[12] ITU-R BT.2022, "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays," International Telecommunication Union, August 2012.

[13] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd - a framework for crowd-based quality evaluation," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 245–248.

[14] ITU-T P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," International Telecommunication Union, July 2012.

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

[16] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *2011 18th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2011, pp. 3097–3100.

[17] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowd workers proven useful: A comparative study of subjective video quality assessment," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.