A new framework for interactive quality assessment with application to light field coding

Irene Viola and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne, Switzerland

ABSTRACT

In recent years, light field has experienced a surge of popularity, mainly due to the recent advances in acquisition and rendering technologies that have made it more accessible to the public. Thanks to image-based rendering techniques, light field contents can be rendered in real time on common 2D screens, allowing virtual navigation through the captured scenes in an interactive fashion. However, this richer representation of the scene poses the problem of reliable quality assessments for light field contents. In particular, while subjective methodologies that enable interaction have already been proposed, no work has been done on assessing how users interact with light field contents. In this paper, we propose a new framework to subjectively assess the quality of light field contents in an interactive manner and simultaneously track users behaviour. The framework is successfully used to perform subjective assessment of two coding solutions. Moreover, statistical analysis performed on the results shows interesting correlation between subjective scores and average interacton time.

Keywords: plenoptic function, light field, subjective evaluation, image coding, user interaction, user behaviour, pattern analysis.

1. INTRODUCTION

The concept of Light Field (LF) as a complete representation of the appereance of a scene was first formalized by Andreī Gershun in his book on radiometric properties of light in 3D space.[1] One of the most popular ways of representing the LF is to describe the radiance along the light rays in a 3D space with constant illumination. It can be achieved using the *plenoptic function*, first introduced in 1991 by Adelson and Bergen.[2] Specifically, the plenoptic function $\mathcal{L}$ describes the intensity of the light rays passing through every possible point in space (V_x, V_y, V_z) at every possible angle (θ, ϕ), wavelength λ, and time t:

$$\mathcal{L} = \mathcal{L}(\theta, \phi, \lambda, t, V_x, V_y, V_z). \tag{1}$$

Considering a 3D region free of occlusions at a single time instance, and bearing in mind the fact that radiance along rays remains constant in a free space, the above 7D plenoptic function can be further simplified into a 4D LF function.[3] The 4D function can be simply parametrized as an intersection of rays with two planes: the uv plane, which describes the rays position in the aperture (object) plane, and the xy plane, describing the rays position in the image plane.

$$\mathcal{L} = \mathcal{L}(u, v, x, y). \tag{2}$$

A digital 4D LF can be obtained by sampling the 4D LF function defined in Equation 2. The digital 4D LF can be seen as a collection of perspective images of the xy plane, each observed from a position on the uv plane. By considering images saved in RGB format, it can be defined as follows:

$$L = L[u, v, x, y, l], \tag{3}$$

Further author information: (Send correspondence to authors) E-mail: {firstname.lastname}@epfl.ch

in which $u = 1, 2, \ldots, U$ and $v = 1, 2, \ldots, V$ represent the indexes of one perspective view in the horizontal and vertical axis, respectively, $x = 1, 2, \ldots, W$ and $y = 1, 2, \ldots, H$ are the pixels comprising each perspective view, and $l = R, G, B$ is the color channel.

Due to the enhanced features and the enriched representation that LF imaging offers, a vast amount of data is created during the acquisition step. Therefore, an efficient way to compress LF images for transmission and storage needs to be designed and implemented. In 2014, the JPEG standardization committee launched a new activity called JPEG Pleno, which aims at creating a standard framework for efficient storage and delivery of LF, point-cloud and holographic content. In particular, the goal of JPEG Pleno is to find an as small as possible number of representation models for these types of content that when necessary also can offer inter-operability with existing solutions, such as legacy JPEG and JPEG 2000 formats. The JPEG Pleno activity for LF contents is currently actively pursuing the definition of a new standard representation and compression algorithm for LF images. Its efforts culminated in a Call for Proposals (CfP) for LF coding solutions, launched jointly with ICIP 2017 Grand Challenge on LF Image Coding*.

Design and validation of new compression solution cannot forgo the importance of reliable assessment of visual quality. In particular, subjective evaluation of visual quality is of fundamental importance when deciding which compression solution should be used. However, quality assessment of LF contents poses new challenges, due to the enriched nature of the content and the possibilities it offers for the rendering step. Two main approaches can be distinguished in LF rendering. The first approach uses the LF information to create a multi-view, 3D rendering of the content, through LF displays or simulators.[4,5] With this approach, the full potential of LF is exploited to create a 3D representation of the scene in front of the user. However, LF displays are not as widely available as 2D displays, mainly due to their cost and their requirements. The second approach uses the LF information to create image-based rendering of LF capabilities.[6] The image-based rendering approach can be used to create several instances of the plenoptic function defined in Equation 2 as 2D images, which can then be displayed on conventional screens. A wide range of possibilities are present in image-based rendering of LF contents. For example, it is possible to combine different perspective views to change the focal plane, in an interactive way. These peculiarities have to be mirrored in the methodology used for subjective quality evaluation.

Assessing the way users engage with LF contents plays a major role on how those methodologies are designed and used. However, user behaviour when engaging with LF content has not yet been studied in details.

In this paper, we propose a new framework for visual quality assessment of LF contents, which allows for interaction with the content and assessment of user experience by tracking user behaviour information. Such information can be subsequently used to further analyze patterns in user interaction. Applications of user interaction information can be found in development of new objective metrics, new subjective methodologies and new perceptual coding algorithms.

The remainder of the paper is organized as follows. Section 2 presents relevant work in LF quality evaluation. Section 3 presents the new methodology in details, whereas Section 4 exposes the validation experiments, and Section 5 details the statistical analysis tools that were used to process the results. Results of the experiments are discussed in Section 6. Section 7 concludes the paper.

2. RELATED WORK

Evaluation of visual quality and user experience plays a fundamental role in designing effective and efficient compression solutions, as well as new rendering techniques. However, only few publications are focused on discussion about objective metrics and subjective methodologies for LF content.

Some publications have been devoted to comparison and evaluation of state of the art standard solutions through objective metrics. Alves et al. assessed the performance of existing still image coding solutions, such as JPEG 2000 and AVC, on lenslet images.[7] The objective evaluation was carried out using PSNR as a full reference metric. Similarly, Vieira et al. compared five different HEVC compatible coding of lenslet images with different data formats,[8] again using PSNR as a full reference metric. Rizkallah et al. reported the impact of

*http://mmspg.epfl.ch/ICIP2017GrandChallenge

compression of LF images on refocusing and extended focus images.[9] They also propose an objective metric to properly assess the amount of compression blur in LF images.

Some preliminary work has been performed on subjectively assess the quality of LF contents on LF displays. Spatial resolution of LF displays has been investigated by Kovacs et al.[10] In particular, the authors investigate how viewing angle affects the perception of spatial resolution, along with the role played by motion parallax. Darukumalli et al. investigate the relationship between zooming levels, region of interest and subjective quality of LF contents, using Absolute Category Rating (ACR) and Degradation Quality Rating (DCR).[11] Kara et al. analyse the impact of angular resolution on the perception of LF content, first in a free movement scenario, and then with fixed observer position, using ACR.[12,13]

LF displays, although commercially available, have not yet seen a widespread success, mainly due to their cost and the requirements for room setup. On the other hand, image-based rendering of LF contents represents a way to engage with LF contents using legacy 2D displays, which are widely available to consumers. However, few publications have been focusing on QoE for image-based rendered LF contents.

In the framework of the Grand Challenge organized at ICME 2016 under a collaboration between Qualinet and JPEG standardization committee, new compression solutions for LF images were evaluated using both objective and subjective quality assessment methodologies.[14] The evaluation was performed on several perspective and refocused images 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). Since the assessment was conducted separately on predefined views, it did not address the issue of evaluating global quality of experience offered by a compressed LF image.

Recently, we proposed a new methodology to evaluate plenoptic contents in an interactive way[15] allowing users to interact with LF images, visualize different views, apply refocusing, and globally evaluate the quality of LF images. The methodology was successfully used to evaluate how different approaches in LF compression can affect the visual quality of the content.[16] Moreover, its validity was tested against a standard passive methodology, which allowed the users to see a pre-recorded animation of the LF content, without any possibility to interact with it.[17]

3. PROPOSED FRAMEWORK

Image-based rendering of LF contents offers the possibility of changing the appereance of the scene as it was captured from the acquisition device: the perspective can be changed, the focal plane can be moved, shrinked or extended, and depth planes can be bypassed, among other visual effects. Such a rich scene representation poses new challenges in assessing the visual quality of the content.

The most precise way of evaluating the visual quality of LF contents would be to render all possible visual effects and rate them separately as 2D images, in order to obtain a single score for any rendering effect that is being tested. A single quality score for the entire LF content can be obtained by averaging the scores obtained from each rendered view. However, averaging the quality of the rendered views does not always give a good indication of the quality of the LF content as a whole. In fact, the final score is highly dependent on the selection of the rendering parameters and on the weights that are assigned to every single rendered view. Moreover, it is clear that this approach is hardly scalable and unfeasible even for a small number of parameters, since the number of stimuli to be tested would notably increase.

An alternative way to evaluate visual quality of LF contents involves a passive approach, where the subjects are presented with a pre-recorded animation displaying different rendered views. Only one comprehensive score is given for the entire presentation. Merging rendered views into a prerecorded presentation helps reducing the complexity of the methodology described above. Moreover, it ensures that all the subjects will have the exact same experience. However, this approach completely disregards the interactive nature of LF contents, reducing it to traditional media contents.

The most natural way of experiencing the capabilities of LF image-based rendering is by enabling interaction with the content in a real-time framework. The possibility of interaction with the acquired content by changing the appearance of the scene has already been proven as a desirable feature in mainstream social media, such as

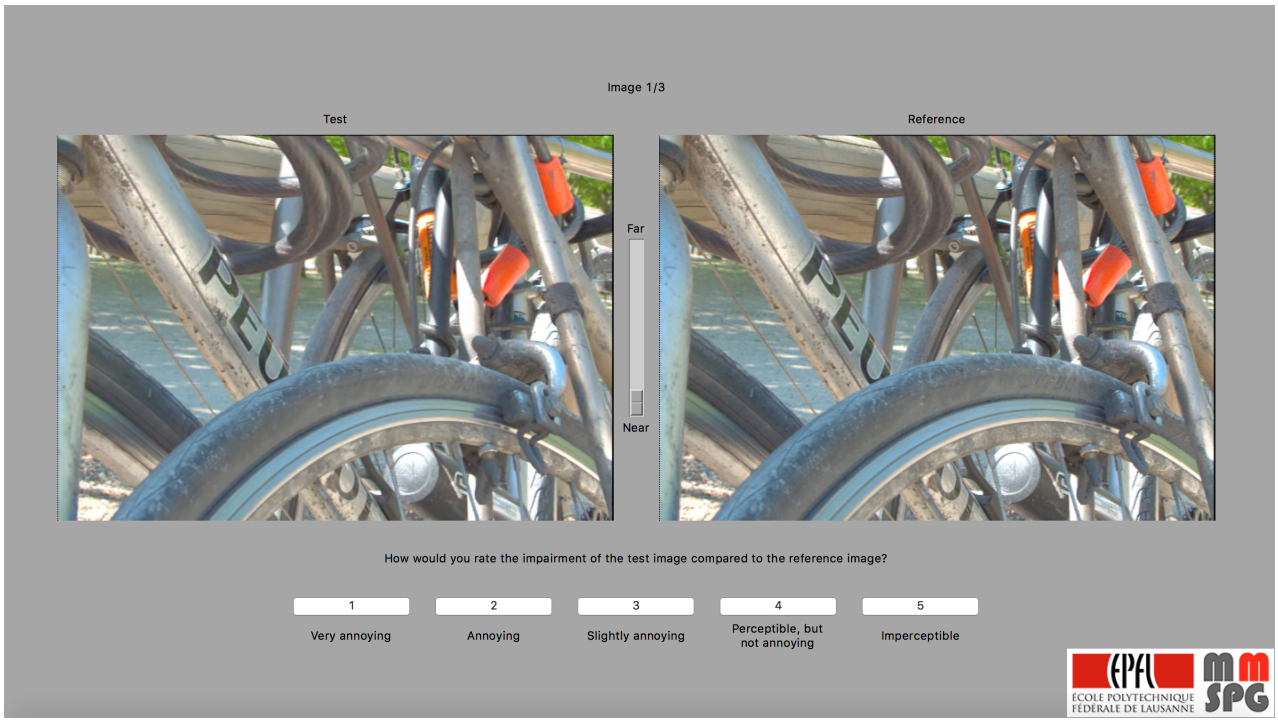


Figure 1: Example of evaluation interface screen.

Instagram, Snapchat and Facebook. A new interactive assessment methodology was already introduced by the authors,[15] and it was successfully used in evaluation of quality of LF contents in presence of compression artefacts.[16] However, it was shown that the interactive methodology had less discriminative power when compared to a passive approach, and leads in general to larger confidence intervals.[17]

In order to retain the discriminative power witnessed in passive approaches, while maintaining the interactive features that define LF contents, a thorough analysis of user behaviour when engaging with such content is needed. Although the subject has been investigated in relation to stereoscopic or LF displays, no work has been presented on analysing patterns in user interaction for image-based rendering.

Three main areas of impact can be identified regarding the analysis of user behaviour:

- **Perceptual coding**. Tracking user behaviour with LF contents leads to statistically accurate knowledge on which rendered views composing the LF content are perceptually meaningful and are more frequently accessed. This information can be used when designing new compression algorithms based on perceptual features.
- **Weighting of objective metrics**. Currently available objective metrics for LF contents, such as those used in ICIP 2017 Grand Challenge on Light Field Coding, give the same weight to all rendered views. A weighted average based on relative frequency of access for every view could be more effective in predicting subjective scores for new LF contents.
- **Design of subjective methodologies**. Interactive subjective methodologies have been shown to be less discriminative than passive subjective methodologies, one of the main reasons being that not all subjects are visualizing strictly the same content. Analysing how users interact with the content will help designing new tests that incorporate user behaviour, thus bridging the gap between interactive and passive subjective methodologies.

The framework proposed in this paper provides a tool to further understand the impact of user behaviour in quality assessment of image-based LF rendering. It consists of a software application for quality assessment of LF contents that enables interaction with the content while keeping track of what the user chooses to visualize. In its current version, the stimuli-comparison method Double Stimulus Impairment Scale (DSIS) is implemented, although the implementation of other methodologies through the framework is straightforward. A graphical interface allows interaction with LF contents in a real-time scenario, by enabling the change of point of view (pespective) and the choice of different focal points (refocus) from a predefined set. Figure 1 shows an example from the framework.

The software takes as input two collections of perspective views in *png* file format, one serving as test and the other serving as reference. The perspective views are then assembled to form the LF content, composed of $U \times V$ images of resolution $W \times H$, according to Equation 3. Additionally, the software can receive as input a set of S images rendered from the LF content at different focal points, which we will refer to as refocused views, and a depth map D that can be used to access the refocused views. Both refocused views and depth map are saved in *png* file format. Test and reference materials must have the same resolution; moreover, they need to be rendered with the same parameters.

The central perspective view from the LF content taken as input is displayed as default for both reference and test materials. By click-and-drag inside the rendered images, the user can change the perspective view, which is rendered in real time. A slider between the two rendered images allows access to the refocused views. Labels on each side of the slider indicate if the content will be refocused on the foreground or on the background. Additionally, the refocused views can be accessed by double clicking on any point of the image. In this case, the depth map is used to map the refocused views to each region in the scene. By clicking and dragging in any point of the rendered image, the user can return to visualize the perspective views. The two contents are rendered simultaneously and they are perfectly synchronized, so the displayed views are rendered with the same parameters. A panel on the bottom of the screen shows the possible scores for the test material. As soon as the user selects one option, the screen updates with the new test material.

The results of the evaluation are saved in one text file. Another text file provided as output records every perspective and refocused view that was accessed by the user, in access order. The start and end times of visualization of each view are recorded, along with the total display time.

4. VALIDATING EXPERIMENT

This section describes the validating experiment in details. More specifically, the creation of the test material and its description are presented as well as the description of the testing environment. Then, a delineation of the test methodology and test plan is provided.

4.1 Dataset preparation and description

For the experiment five LF contents, acquired with a Lytro Illum camera, were chosen from EPFL LF image dataset.[18] More specifically, contents *Bikes*, *Danger_de_Mort Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* were used in our experiments.

Each 10bit raw lenslet image was devignetted, demosaiced, and transformed into an LF data structure of perspective views using the Light Field toolbox v0.4.[19,20] A total of 15×15 perspective views were created from the lenslet image, each having a resolution of 625×434 pixels. The perspective views were subsequentially saved in *ppm* file format, with 10 bits per color channel, to serve as reference.

Two codecs were adapted for compression of LF evaluated in the test. Both codecs perform the compression on the perspective views, which were preemptively ordered in a pseudo-temporal sequence. To be used as input for the compression algorithms, the perspective images were padded with black pixels, converted to YCbCr format and downsampled from 444 to 422, 10-bit depth. Then, they were arranged in a pseudo-temporal arrangement (see Figure 2) and saved in *yuv* file format.

The first codec that was used to compress the pseudo-temporal sequence consisted in HEVC Main10 profile. The software x265 was used to compress the sequence[†]. The full command line used can be found in Table 1.

[†]https://www.videolan.org/developers/x265.html

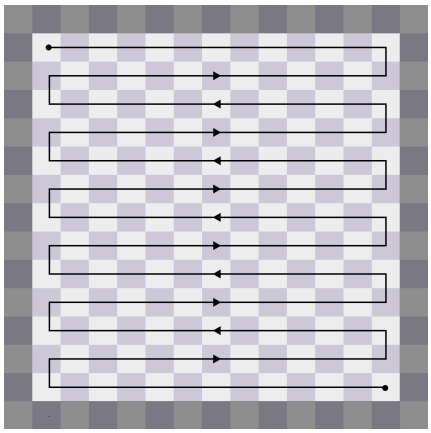


Figure 2: Order of perspective views for pseudo-temporal sequence used for coding.

Table 1: Selected settings for x265 Main10 coder.

--input < *Input* > --input-depth 10 --input-csp i422 --fps 30 --input-res < *Width* > × < *Height* > --output < *Output* > --output-depth 10 --profile main422-10 --crf < *QP* >

Table 2: QP chosen to encode all contents with HEVC.

Content	**R1**	**R2**	**R3**	**R4**
Bikes	13	24	33	44
Danger_de_Mort	15	26	35	43
Stone_Pillars_Outside	14	23	30	40
Fountain_&_Vincent_2	14	24	32	43
Friends_1	12	21	29	40

Table 3: Selected settings for VP9 coder.

--i422 --input-bit-depth=10 --profile=3 -w < *Width* > -h < *Height* > --target-bitrate=< *bitrate*> --cq-level=0 --bit-depth=10 --codec=vp9 --fps=30000/1000 --best -o < *Output* > < *Input* >

The Quantization Parameters (QP) were chosen to match the desired compression ratios. Table 2 summarizes the values of different QP used in the test.

As a second codec, VP9 was used to compress the pseudo-temporal sequence[‡]. The full command line used can be found in Table 3. The target bitrate was chosen to match the corresponding compression ratios defined below.

The test LF content was displayed together with the uncompressed reference in a side-by-side fashion, using the proposed framework. Due to distorsions naturally occurring in lenslet-based LF content, some of the perspective views were deemed not suitable for visualization, since they would negatively bias subjects. Hence, only the central 9×9 perspective views out of the 15×15 views were selected for the test. Both reference and test contents were converted from *ppm* file format in 10 bits to *png* file format in 8 bits, due to limitations of the display and the software.

[‡]https://www.webmproject.org/vp9/

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

Figure 3: Central perspective view from each content used in the test.

Table 4: Values of refocusing slope 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
Danger_de_Mort	-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

For each stimulus, the central perspective view from the LF data structure was initially displayed. By clicking inside either test or reference displayed image and dragging the mouse, the other perspective views from the data structure were accessed and displayed. Each image was displayed in its native resolution of 625×434 pixels. Additionally, eleven refocused images of the central perspective view were created for each content, using a modified version of the toolbox function *LFFiltShiftSum*. The function shifts all the images in the stack according to a parameter, called slope, and performs a sum of the shifted images to obtain a single image that is refocused on a specific plane, which depends on the value of the slope. The number of images to be shifted and consequently summed defines the depth of field. Summing all 15×15 images creates the smallest depth of field, in which only one specific plane in the image is in focus. On the other hand, taking just the central image, which is equivalent to summing just 1×1 images, brings all the objects in focus (largest depth of field). For the test, it was chosen to sum images from index 3 to index 13 (11×11 images) to have a depth of field that still shows the effects of refocusing. The values of the slopes are summarized in Table 4. The refocused images were accessible through a slider shown between test and reference. Additionally, users could access the refocused images by double clicking on the point of the image they wished to see in focus. The slopes were selected so as to assure gradual transition between refocusing on the foreground and on the background with respect to semantically relevant objects in each content.

The codecs were evaluated on four bitrates, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, $R4 = 0.005$ bpp. The compression ratios are computed as ratios between the size of the uncompressed raw images in 10-bit precision ($5368 \times 7728 \times 10$ bits $= 414839040$ bits $= 10$ bpp) and the size of the compressed bitstream.

4.2 Testing environment

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.[21] A Samsung SyncMaster 2443 24-inch monitor with native resolution of 1920×1200 pixels was used for the test. The monitor was calibrated using an i1Display Pro color calibration device according to the following profile: sRGB Gamut, D65 white point, 120cd/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

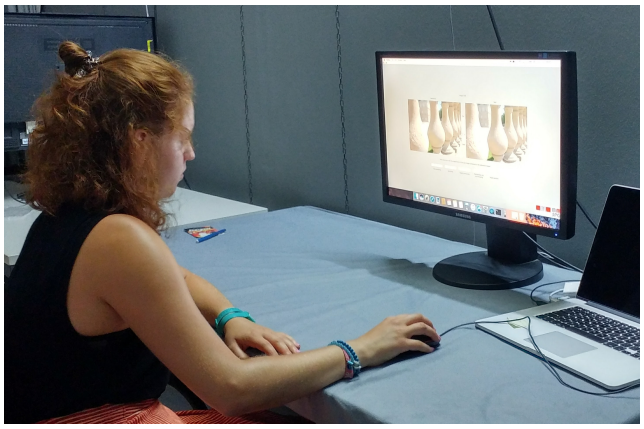


Figure 4: Testing environment.

height of the displayed content, conforming to requirements in ITU-R Recommendation BT.2022.[22] A picture of the MMSPG test laboratory serving as an evaluation environment is shown in Figure 4.

4.3 Test methodology and planning

The selected methodology was based on DSIS.[21] The participants were asked to interact with the LF contents and to rate the level of impairment of the test LF content with respect to the reference, on a scale from 1 (Very annoying) to 5 (Imperceptible). Each LF content was presented together with the uncompressed reference in a side-by-side fashion, in its native resolution of 625×434 pixels. The position of the reference was fixed for each experiment, and the participants were made aware of its location on the screen (either left or right).

Before the experiments, a training session was organized to allow participants to get familiar with artifacts and distorsions in the test images. Four training samples were manually selected by expert viewers. In order not to influence the results, the training samples were created by compressing other contents on various bitrates. The content used for the training was chosen from the same LF database used for the test images.[18] The training samples were presented along with the uncompressed reference, exactly as they were shown in the test.

The test samples were randomly distributed among subjects. The same content was never shown consecutively. Before the test, two dummy samples were inserted to ease the participants into the task. The resulting scores from dummy stimuli were not included in the results.

A total of 23 subjects (11 males and 12 females) participated in the experiment, for a total of 23 scores per stimulus. Subjects were between 18 and 35 years old, with an average of 22.27 and a median of 22.05 years of age. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts.

5. DATA PROCESSING AND STATISTICAL ANALYSIS

This section describes how data was processed to obtained the results presented in next section. Specifically, subsection 5.1 details how subjective scores were processed and analyzed, subsection 5.2 enlists the pre-processing and aggregation of user tracking data, while subsection 5.3 presents the statistical analysis and cross-correlation of the results.

5.1 Subjective scores analysis

Outlier detection was performed according to the guidelines defined in ITU-R recommendation BT.500-13.[21] One outlier was detected and the relative scores were discarded, thus leading to 22 scores per stimulus. The Mean Opinion Score (MOS) was computed for each coding condition j (i.e., each content, codec and compression ratio) as follows:

$$MOS_j = \frac{1}{N}\sum_{i=1}^{N} m_{i,j}, \tag{4}$$

where N is the number of participants and m_{ij} is the score for stimulus j by participant i. The corresponding 95% confidence intervals were computed. To determine whether the results yield statistical significance, a one-sided Welch's test at 5% significance level was performed on the scores, with the following hypotheses:

$$H0 : MOS_A \leq MOS_B$$
$$H1 : MOS_A > MOS_B,$$

in which A and B are the codecs that are being compared. The test was performed for each compression ratio and for each content. If the null hypothesis were to be rejected, then it could be concluded that codec A performed better than codec B for the given content and compression ratio at a 5% significance level.

5.2 Tracking information analysis

The total number of seconds spent on each perspective and refocused view are aggregated for each stimulus and for each subjects in matrices $P_{i,j}$ and $R_{i,j}$, respectively:

$$P_{i,j} = \begin{pmatrix} p_{1,1,i,j} & \cdots & p_{1,v,i,j} & \cdots & p_{1,V,i,j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{u,1,i,j} & \cdots & p_{u,v,i,j} & \cdots & p_{u,V,i,j} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{U,1,i,j} & \cdots & p_{U,v,i,j} & \cdots & p_{U,V,i,j} \end{pmatrix}, R_{i,j} = \begin{pmatrix} r_{1,i,j} \\ \vdots \\ r_{s,i,j} \\ \vdots \\ r_{S,i,j} \end{pmatrix}, \tag{5}$$

where $u = 1, 2, \ldots, U$ and $v = 1, 2, \ldots, V$ are the indexes of each perspective view, $s = 1, 2, \ldots, S$ is the index of each refocused view, $i = 1, 2, \ldots, N$ indicates the subject and $j = 1, 2, \ldots, M$ indicates the stimulus.

The results were then aggregated to get the total number of seconds each subject spent on each stimulus:

$$\bar{P}_{i,j} = \sum_{u=1}^{U}\sum_{v=1}^{V} p_{u,v,i,j}, \tag{6}$$

$$\bar{R}_{i,j} = \sum_{s=1}^{S} r_{s,i,j}, \tag{7}$$

$$\bar{T}_{i,j} = \bar{P}_{i,j} + \bar{R}_{i,j}. \tag{8}$$

To get the general trend for each stimulus, the mean was computed across all subjects:

$$\hat{P}_j = \frac{1}{N}\sum_{i=1}^{N} \bar{P}_{i,j}, \tag{9}$$

$$\hat{R}_j = \frac{1}{N}\sum_{i=1}^{N} \bar{R}_{i,j}, \tag{10}$$

$$\hat{T}_j = \frac{1}{N}\sum_{i=1}^{N} \bar{T}_{i,j}. \tag{11}$$

5.3 Correlation and validation analysis

Statistical analysis was performed on the subjective scores and the results obtained from the tracking of user behaviour, to see whether the results obtained presented some correlation. In particular, statistical analysis was performed between MOS_j, which was used as ground truth, and $\hat{P}_j$, $\hat{R}_j$ and $\hat{T}_j$, for a total of three comparisons. For simplicity, from now on we will refer to $\hat{P}_j$, $\hat{R}_j$ and $\hat{T}_j$ as tracking values.

Following the ITU-T Recommendation P.1401,[23] several fittings were applied to the tracking values. In particular, first order and third order fittings were used to compare the 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.

In order to understand whether the tracking values could effectively be used as predictors for MOS values, estimation and classification errors were computed. A multiple comparison test was performed at a 5% significance level on the raw scores, to determine, for each stimulus, whether the MOS values and the tracking values were significantly different, and the percentage of correct estimation, underestimation and overestimation were computed. Underestimation occurs when the MOS value predicted from the tracking values is significantly lower than the true MOS value. Overestimation, on the other hand, occurs when the predicted MOS value is significantly higher than the true value.

The classification errors were computed using the same multiple comparison test to see if the results lead, for each pair of stimuli, to the same conclusions.[24] In this case, three types of errors can be distinguished: false ranking, false differentiation and false tie. False ranking, the most offensive error, occurs when the ground truth says that situation j_1 is better than situation j_2, whereas the predicted MOS obtained from tracking values say the opposite. False differentiation occurs when the true MOS values say that situation j_1 and j_2 are the same, whereas the prediction from tracking results says they are different. False tie occurs when the true MOS scores say two situations are different, whereas the predicted MOS scores say they are the same.

6. RESULTS AND DISCUSSION

This section describes and discusses the results obtained in the evaluation campaign. More specifically, subjective evaluation results are introduced in section 6.1. Section 6.2 presents the insights provided by tracking of user behaviour, while section 6.3 details the correlation and validation results.

6.1 Subjective evaluation results

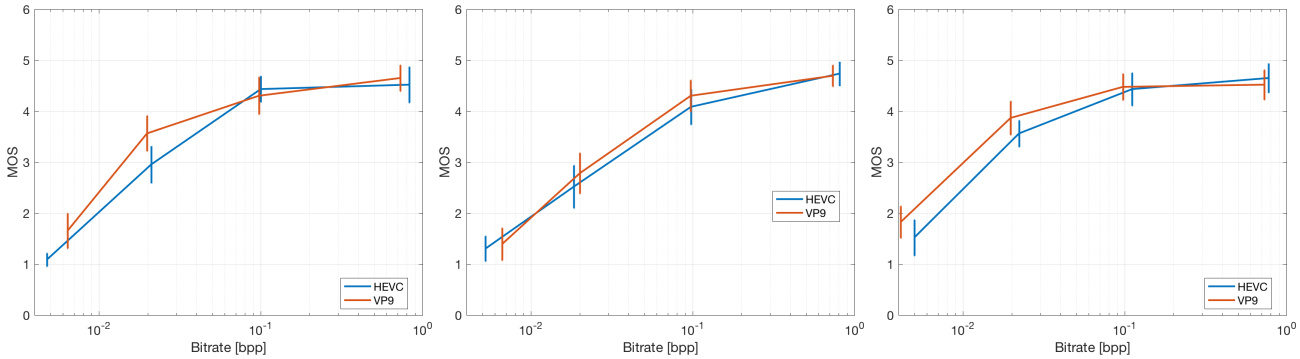
Figure 5 shows the MOS against bitrate for all the contents under test, with respective confidence intervals. It can be observed that while the codecs have very similar performance on compression ratio $R1$ and $R2$, some difference can be observed for compression ratios $R3$ and $R4$, where VP9 outperforms HEVC in some of the contents.

The observation is confirmed by the results obtained in Welch's test, summarized in Table 5. HEVC is never significantly better than VP9. For compression ratio $R2$, the two codecs are statistically equivalent, whereas VP9 outperforms HEVC on one out of five contents for compression ratio $R1$, two out of five contents for compression ratio $R3$, and three out of five contents for compression ratio $R4$.

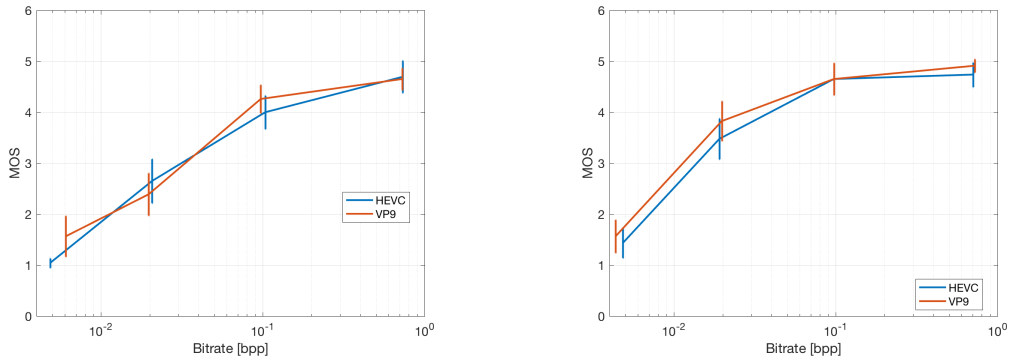
6.2 User tracking results

Figure 6 shows the total interaction time $\bar{T}_{i,j}$ for each subject i and each stimulus j. It is noticeable that some users engaged longer with the contents, while others spent on average very little time interacting with the contents (see for example column 6 and 18).

However, a clear trend can be observed among the stimuli. Users seem to spend less time with contents compressed at compression ratio $R4$, as it is visible from the horizontal dark lines in Figure 6. In particular, darker lines are present for contents compressed with HEVC at the lowest compression ratio.



(a) Bikes (b) Danger_de_Mort (c) Stone_Pillars_Outside



(d) Fountain_&_Vincent_2 (e) Friends_1

Figure 5: MOS vs bitrate for different contents. The bitrate is shown in logarithmic scale to improve readability.

Table 5: Number of contents for which the null hypothesis was rejected, for each compression ratio.

Codec	**R1**	**R2**	**R3**	**R4**
HEVC	0	0	0	0
VP9	1	0	2	3

In order to see whether the order of presentation of the stimuli for each subject had any influence on total interaction time $\bar{T}_{i,j}$, the total interaction time was displayed for every subject following the presentation order (see Figure 7). No definite trend can be observed. Hence, it can be concluded that the presentation order had no influence on the total time the users spent interacting with the content. Subjects' boredom and fatigue, along with repetitiveness of the contents, did not have a definite impact on the total time they spent engaging with the stimuli.

Figures 8 and 9 show the average interaction time $\hat{P}_j$, $\hat{R}_j$ and $\hat{T}_j$, divided by content and by compression ratio, for codec HEVC and VP9, respectively. The results show the trend already observed in Figure 6: on average, users tend to interact more with higher bitrates (compression ratios $R1$ and $R2$), whereas for lower bitrates they tend to interact less (compression ratio $R4$). The trend is visible for all type of interactions and for both codecs, although on average people tend to spend more time on codec VP9 for compression ratio $R4$ than they do on codec HEVC.

In general, results are more polarized for codec HEVC: users tend to interact more with content compressed with HEVC at higher bitrates with respect to the VP9 counterpart, but they also tend to interact less with content compressed with HEVC at lowest bitrate than they do with the same content compressed with VP9. The average interaction time for codec VP9 is more evenly distributed, although a bitrate-depended trend is still

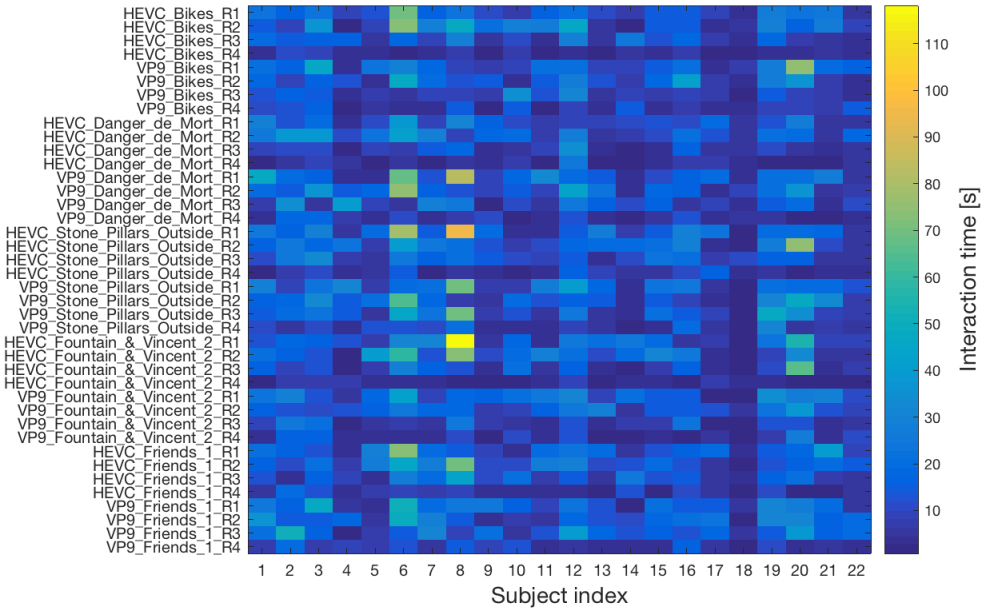


Figure 6: Total interaction time $\bar{T}_{i,j}$ (in seconds) vs stimuli vs subjects.

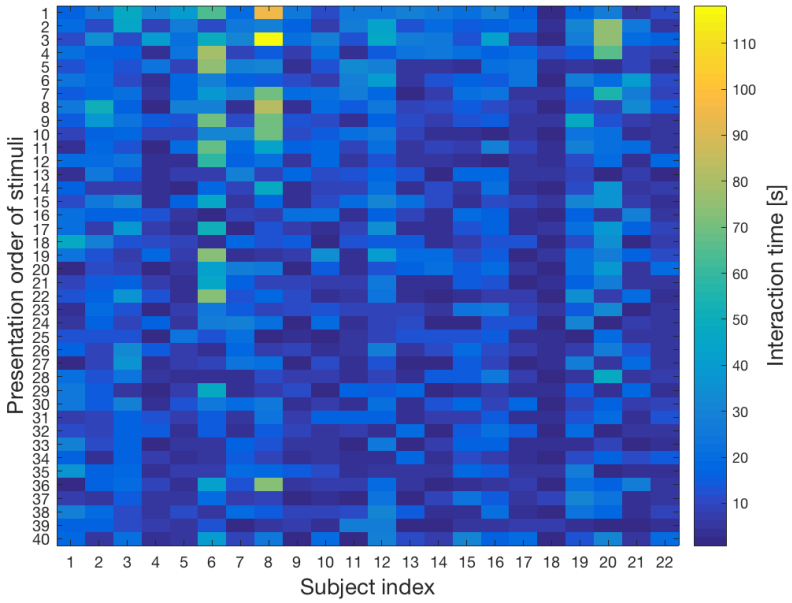
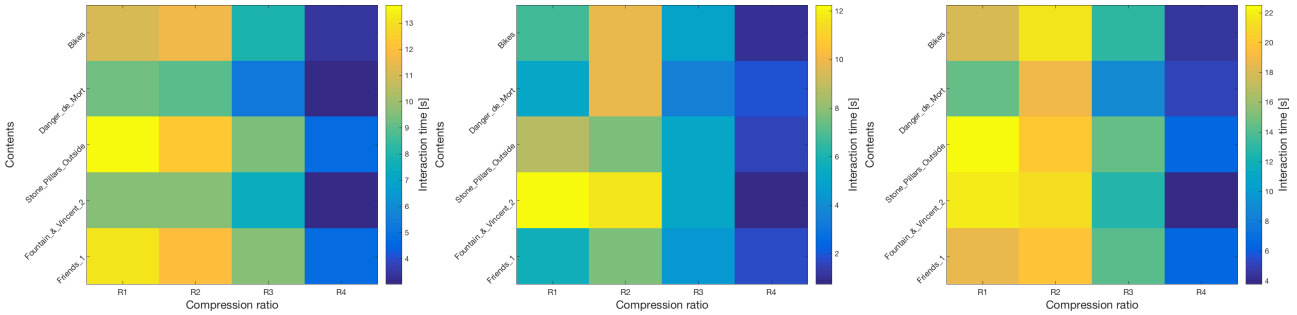


Figure 7: Total interaction time $\bar{T}_{i,j}$ (in seconds) vs order of presentation of the stimuli for each subject.

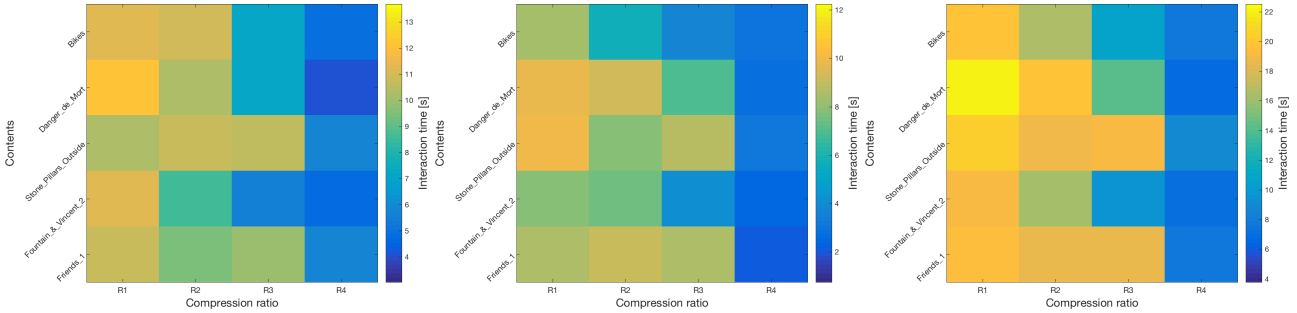
clearly visible (see Figure 9).

The average interaction time with perspective views and with refocused views, for both codecs, follows an alternated trend: if users on average spent more time interacting with perspective views for a certain content, they would consequentially spend less time interacting with refocused views. The phenomenon is particularly evident for content *Fountain_&_Vincent_2*. When compressed with codec HEVC at compression ratio $R1$, the content saw an increase in interaction with refocused views, to the detriment of average time spent interacting



(a) Interaction time for perspective views. (b) Interaction time for refocused views. (c) Interaction time for all views.

Figure 8: Average interaction time or perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), divided by content and by compression ratio, for codec HEVC.



(a) Interaction time for perspective views. (b) Interaction time for refocused views. (c) Interaction time for all views.

Figure 9: Average interaction time or perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), divided by content and by compression ratio, for codec VP9.

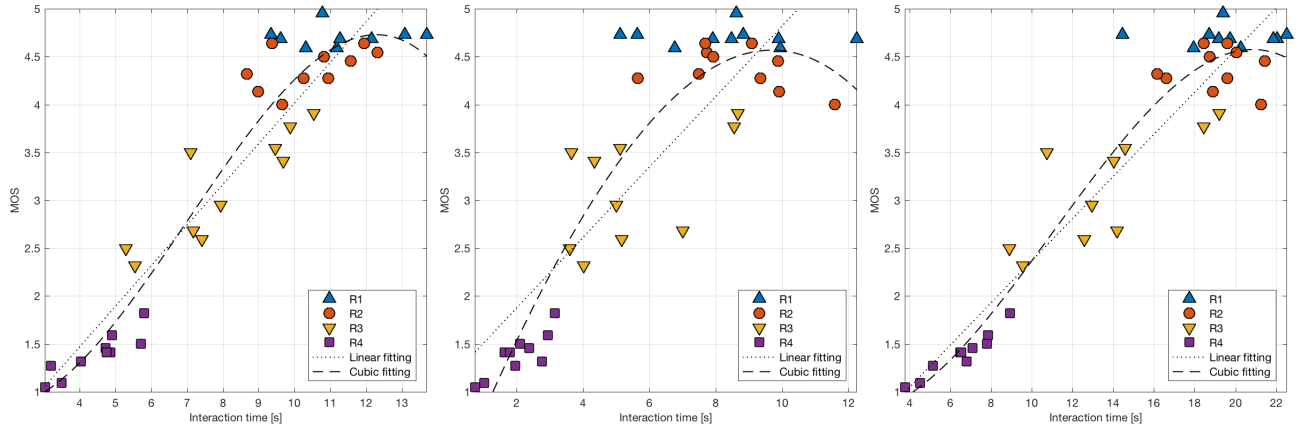
with perspective views. The opposite behaviour can be observed for codec VP9. However, the total interaction time is quite similar between the two codecs (21.86 against 19.19 seconds).

A trend is also visible regarding the contents, at least when compressed using HEVC (see Figure 8). The average interaction time with perspective views was generally higher for contents *Stone_Pillars_Outside* and *Friends_1*, followed by *Bikes*. On the other hand, *Fountain_&_Vincent_2* and *Danger_de_Mort* were, on average, the contents for which the users engaged the least when they needed to interact with perspective views. Interestingly enough, when analysing interaction with refocused views, users engaged more with *Fountain_&_Vincent_2* and *Danger_de_Mort* than they did with the other three contents. As a result, the total average interaction time with all views shows that no visible trend is present for different contents.

6.3 Correlation and validation

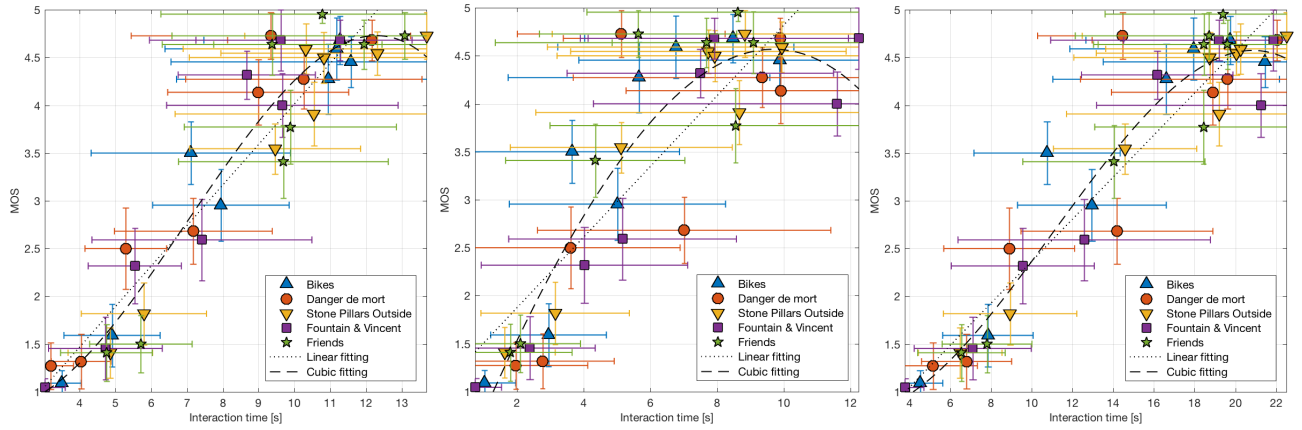
Figure 10 shows the scatter plots comparing the MOS values to the average interaction time for perspective views $\hat{P}_j$, refocused views $\hat{R}_j$, and all views $\hat{T}_j$. To improve visualization, the points were colored based on compression ratio. Figure 11 shows the same scatter plots with respective CIs. In this case, points were colored based on the content. Linear and cubic regressions are shown for all comparisons. Table 6 shows the performance indexes computed on the data. The indexes are computed on data pairs $[\widehat{X}, MOS]$, in which MOS is the ground truth, and $\widehat{X} = \widehat{P}, \widehat{R}, \widehat{T}$ are the average interaction time results after linear and cubic fitting.

Results from linear and cubic fitting confirm the trend already observed in the previous section. In particular, lower MOS scores are associated with less interaction time on average, whereas longer interaction time is associ-



(a) Interaction time for perspective views vs MOS. (b) Interaction time for refocused views vs MOS. (c) Interaction time for all views vs MOS.

Figure 10: Average interaction time for perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), vs MOS. The points are differentiated by compression ratio.



(a) Interaction time for perspective views vs MOS. (b) Interaction time for refocused views vs MOS. (c) Interaction time for all views vs MOS.

Figure 11: Average interaction time for perspective views $\hat{P}_j$ (a), refocused views $\hat{R}_j$ (b), and all views $\hat{T}_j$ (c), vs MOS, with respective CIs. The points are differentiated by content.

ated with higher MOS scores, for $\hat{P}_j$, $\hat{R}_j$, and $\hat{T}_j$. Results are further confirmed by values obtained performing PCC and SRCC, which show a strong linear correlation.

CIs associated with average interaction time also show that greater variations can be observed for higher MOS scores (see Figure 11). As MOS scores decrease, the CIs tend to be smaller as well. Larger CIs can be observed for $\hat{R}_j$ with respect to $\hat{P}_j$ and $\hat{T}_j$.

Although $\widehat{R}$ displays linear correlation with MOS scores, accuracy and consistency are quite low (OR = 82.50% and RMSE = 0.7048 for linear fitting). Indeed, the scatter plot presents a less definite trend with respect to results obtained by $\hat{P}$, especially for compression ratios $R1$ and $R2$ (see Figure 10 (a) and (b)). However, adding the results of the interaction with refocused views to the results of the interaction with perspective views helps with both accuracy and consistency (see Figure 10 (c) and Table 6).

Results from multiple comparison test show that correct estimation is achieved in 100% of cases with linear fitting, and an under estimation of 2.50% is observable when using cubic fitting. Moreover, false ranking, the

most offensive error, is never present. Correct decision is achieved on more than 95% of the cases when applying cubic fitting. In general, $\widehat{T}$ has slightly better predictive power than $\widehat{P}$ (i.e., using only perspective views as opposed to using also refocused views). It also achieves better consistency and accurancy. It can be concluded that, while the average time spent on refocused views alone cannot be used as a predictor for MOS scores, adding it to the average time spent on perspective views improves the correlation results.

Table 6: Performance indexes.

	$[\widehat{P}, MOS]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Correct Decision**	**False Diff.**	**False Tie**
Linear fitting	0.9408	0.8704	0.4596	45.00%	100%	0.00%	87.69%	1.41%	10.90%
Cubic fitting	0.9613	0.8511	0.3736	30.00%	97.50%	2.50%	95.38%	4.62%	0.00%
	$[\widehat{R}, MOS]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Correct Decision**	**False Diff.**	**False Tie**
Linear fitting	0.8543	0.7677	0.7048	82.50%	100%	0.00%	89.74%	1.67%	8.59%
Cubic fitting	0.9161	0.7667	0.5436	57.50%	97.50%	2.50%	97.05%	2.95%	0.00%
	$[\widehat{T}, MOS]$								
	PCC	**SRCC**	**RMSE**	**OR**	**Correct Est.**	**Under Est.**	**Correct Decision**	**False Diff.**	**False Tie**
Linear fitting	0.9462	0.8568	0.4387	45.00%	100%	0.00%	90.51%	2.82%	6.67%
Cubic fitting	0.9605	0.8255	0.3774	22.50%	97.50%	2.50%	96.15%	3.85%	0.00%

Further validation is needed to confirm the predictive power of average interaction time for subjective scores. However, using average interaction time as a predictor for MOS values lays the basis for an implicit quality assessment methodology. One can envision such methodology could be extremely useful in a near future, when plenoptic content will be available on social media and tracking data can be collected anonymously from the natural interaction users have with the content. The tracking data can then be used to predict the quality of the content the users are engaging with, without asking for an explicit score.

7. CONCLUSION

In this paper we proposed a new framework for tracking user behaviour while performing subjective quality assessment of light field contents. We presented the new framework in details, and we provide an implementation in form of a software interface. We also presented and discussed results obtained by using the proposed implementation. In particular, we presented the results of the subjective assessment and the aggregated results of user tracking. Then, we performed correlation between the two to assess the predictive power of average interaction time for subjective scores. Results showed that the proposed methodology can be successfully used to assess the visual quality of light field content. Moreover, we show that the average interaction time can be used to predict the subjective score of light field contents.

The paper lays the basis of an implicit quality assessment method for light field contents. Further analysis is needed to prove if the average interaction time can effectively be used instead of explicit quality scores to assess the visual quality of light field contents.

A *python* implementation of the proposed framework can be found at the following link: `https://github.com/mmspg/light-field-tracking`.

ACKNOWLEDGMENTS

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). The authors would like to thank Tanguy Albrici for his help bringing the software into life, Alessandro Ebrahimi for his valuable help running the validating experiment, and Evangelos Alexiou and Anne-Flore Perrin for useful comments and fruitful discussions.

REFERENCES

[1] Gershun, A., "The light field," *Journal of Mathematics and Physics* **18**(1), 51–151 (1939).

[2] Adelson, E. H. and Bergen, J. R., "The plenoptic function and the elements of early vision," (1991).

[3] Levoy, M. and Hanrahan, P., "Light field rendering," in [*Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*], 31–42, ACM (1996).

[4] Balogh, T., "The holovizio system," in [*Electronic Imaging 2006*], 60550U–60550U, International Society for Optics and Photonics (2006).

[5] Matsubara, R., Alpaslan, Z. Y., and El-Ghoroury, H. S., "Light field display simulation for light field quality assessment," in [*SPIE/IS&T Electronic Imaging*], 93910G–93910G, International Society for Optics and Photonics (2015).

[6] McMillan, L. and Bishop, G., "Plenoptic modeling: An image-based rendering system," in [*Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*], 39–46, ACM (1995).

[7] Alves, G., Pereira, F., and da Silva, E. A., "Light field imaging coding: Performance assessment methodology and standards benchmarking," in [*2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*], 1–6, IEEE (2016).

[8] Vieira, A., Duarte, H., Perra, C., Tavora, L., and Assuncao, P., "Data formats for high efficiency coding of lytro-illum light fields," in [*2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*], 494–497, IEEE (2015).

[9] Rizkallah, M., Maugey, T., Yaacoub, C., and Guillemot, C., "Impact of light field compression on refocused and extended focus images," in [*2016 24th European Signal Processing Conference (EUSIPCO)*], (2016).

[10] Kovács, P. T., Lackner, K., Barsi, A., Balázs, Á., Boev, A., Bregović, R., and Gotchev, A., "Measurement of perceived spatial resolution in 3d light-field displays," in [*2014 IEEE International Conference on Image Processing (ICIP)*], 768–772, IEEE (2014).

[11] Darukumalli, S., Kara, P. A., Barsi, A., Martini, M. G., and Balogh, T., "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).

[12] Kara, P. A., Martini, M. G., Kovacs, P., Imre, S., Barsi, A., Lackner, K., Balogh, T., 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).

[13] Kara, P. A., Cserkaszky, A., Darukumalli, S., Barsi, A., and Martini, M. G., "On the edge of the seat: Reduced angular resolution of a light field cinema with fixed observer positions," in [*9th International Conference on Quality of Multimedia Experience (QoMEX)*], 1–6, IEEE (2017).

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

[15] Viola, I., Řeřábek, M., and Ebrahimi, T., "A new approach to subjectively assess quality of plenoptic content," in [*SPIE Optical Engineering+ Applications*], 99710X–99710X, International Society for Optics and Photonics (2016).

[16] Viola, I., Řeřábek, M., and Ebrahimi, T., "Comparison and evaluation of light field coding approaches," *IEEE Journal of selected topics in signal processing* (2017).

[17] Viola, I., Řeřábek, M., and Ebrahimi, T., "Impact of interactivity on the assessment of quality of experience for light field content," in [*9th International Conference on Quality of Multimedia Experience (QoMEX)*], (2017).

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

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

[20] Dansereau, D. G., Pizarro, O., and Williams, S. B., "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics (TOG)* **34** (Feb. 2015).

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

[22] 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).

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

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