Comparison of Interactive Subjective Methodologies for Light Field Quality Evaluation

Irene Viola and Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Email: firstname.lastname@epfl.ch

***Abstract*—The recent advances in light field acquisition and rendering technologies have attracted a lot of interest from the scientific community. Due to their large amount of data, efficient compression of light field content is of paramount importance for storage and delivery. Quality evaluation plays a major role in assessing the impact of compression on visual perception. In particular, subjective methodologies for light field quality assessment must be carefully designed to ensure reliable results. In this paper, we present and compare two different methodologies to evaluate visual quality of light field contents. Both methodologies allow users to interact with the content and to freely decide which viewpoints to visualize. However, in the second methodology a brief animation of the available viewpoints is presented prior to interaction in order to ensure the same experience for all subjects. The time and patterns of interaction of both methods are compared and analyzed through a rigorous analysis. Conclusions provide useful insights for selecting the most appropriate light field evaluation methodology.**

I. Introduction

The feature-rich representation offered by Light Field (LF) photography promises new ways of interaction with real-life scenarios in an immersive environment. Recent innovations in acquisition and rendering of LF contents have fueled the interest of the scientific community, due to the challenges that this new representation brings. One of those challenges consists in the large volume of data generated in the acquisition process, which is demanding in terms of storage and transmission. Thus, new compression solutions must be designed to efficiently reduce the amount of data while preserving both visual quality and immersive features. At the same time, new subjective assessment methodologies must be designed and thouroughly tested to evaluate the impact of compression, representation, and rendering models on perceptual quality and user experience.

Several studies have been devoted to subjective quality evaluation of LF contents. Paudyal et al. investigate the impact of watermarking on visual quality of LF contents, and especially the relationship between watermark strength and visual quality, using Absolute Category Rating (ACR) [1]. Darukumalli et al. examine the quality of experience associated to LF displays in relation to zooming levels and regions of interest, using ACR and Degradation Category Rating (DCR) [2]. Kara et al. analyse the correlation between spatial and angular resolution, and how reducing spatial resolution can improve parallax perception [3]. In their previous work, the authors have evaluated several coding approaches and their impact on visual quality, using two different methodologies [4]. They have also compared passive and interactive methodologies, focusing on the impact of interaction on the collected scores [5]. A preliminary study on interaction trends has also been presented [6].

The enriched rendering made possible by LF imaging is most easily explored by interacting with the contents, for example by changing perspective or applying digital refocusing. However, it has been shown that interaction may lead to an increase in the variance of the collected scores, due to the variation between user experiences [5]. Interaction still remains a valuable feature in assessing the visual quality of LF images, since it represents the most natural way of consumption, letting users engage with the content. In such realistic scenario, the analysis of user behavior can be further considered to improve perceptual coding and objective metrics, among others.

In this paper, we combine passive and interactive methodologies to ensure the same visualization experience for all users, while still enabling interaction with LF contents. We compare this visualization and interaction approach to a purely interactive setup, in which no passive-like animation is presented to the subjects. Correlation is computed between the obtained scores to see whether the results are statistically different. Statistical analysis is carried out on the time of interaction associated with the two approaches to examine how user engagement is affected by the testing setup.

The remainder of the paper is organized as follows. Details on how the experiment was designed and carried out, as well as how the scores were processed and analysed, are presented in section II. Results from the comparison are discussed in section III, and conclusions are drawn in section IV.

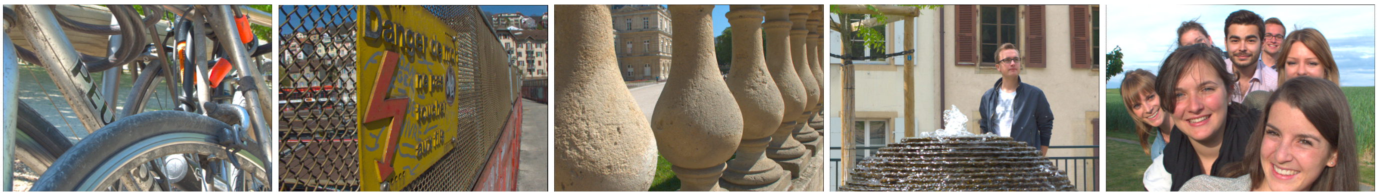
II. Experimental test

A. Data preparation

Five LF contents in 10-bit raw lenslet format were selected from a publicly available LF image dataset, namely *Bikes*, *Danger_de_Mort*, *Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* [7]. The central perspective view from each content is shown in Figure 1.

Each lenslet image was devignetted, demosaiced, and transformed into a stack of perspective views using the Light

ISBN 978-90-827970-1-5 © EURASIP 2018



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

Fig. 1: Central perspective view from each content used in the test.

TABLE I: 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

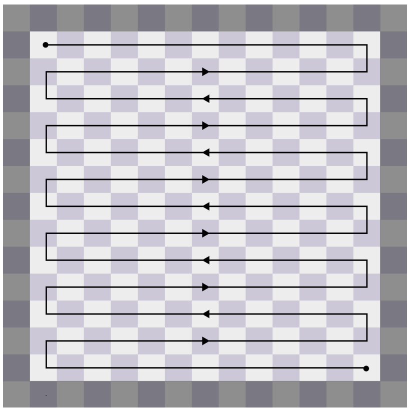


Fig. 2: Order of perspective views for pseudo-temporal sequence used for coding.

Field toolbox v0.4 [8], [9]. From the lenslet image, 15×15 perspective views were generated, each having a resolution of 625×434 pixels. To serve as input for the compression algorithms, the perspective views were padded with black pixels, converted to YCbCr format and downsampled from 444 to 422, 10-bit depth. They were then arranged in a pseudo-temporal arrangement following a serpentine order, as depicted in Figure 2, and saved in *yuv* file format.

Two codecs were adapted for compression of LF. The first selected codec was HEVC. The contents were encoded using the software implementation x265[1], with the Main10 profile. The Quantization Parameters (QP) were selected to match the desired compression ratios. VP9 was used as second codec to compress the pseudo-temporal sequence[2]. The target bitrate was chosen to match the corresponding compression ratios defined below. Details about the software specifications, and QP and target bitrate selection can be found in [6].

The performance of the codecs was evaluated on four bitrates, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, $R4 = 0.005$ bpp. The bitrates are computed by dividing the size of the compressed bitstream over the size of the uncompressed raw images (5368×7728 pixels).

B. Subjective Methodology

To perform the subjective visual quality assessment, the Double Stimulus Impairment Scale (DSIS) methodology with 5-point grading scale (5: *Imperceptible*, 4: *Perceptible but not annoying*, 3: *Slightly annoying*, 2: *Annoying*, 1: *Very annoying*) was selected, based on ITU-R Recommendation BT.500-13 [10].

The test stimuli were displayed side-by-side with the uncompressed reference, using the framework proposed in [6]. Due to distorsions naturally occurring in lenslet-based LF content, some of the perspective views were excluded from the test, since they would negatively bias subjects. Hence, only the central 9×9 perspective views out of the 15×15 views were visualized in the test. Both reference and test contents were converted to png file format in 8 bits, due to limitations of the display and the software. Each image was displayed in its native resolution of 625×434 pixels. Eleven refocused images of the central perspective view were additionally created for each content, using a modified version of the toolbox function *LFFiltShiftSum* that uses the central 11×11 perspective views to create a larger depth of field. The function uses a parameter, called slope, to define the plane where the refocusing will be applied. 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 values of the slopes are summarized in Table I.

The experiment was divided in two sessions, corresponding to two different visualization and interaction approaches. In the first session, the participants could directly interact with the perspective and refocused views. For each stimulus, the central perspective view of the test and reference content was initially displayed, in a side-by-side fashion. By clicking inside

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

[2] https://www.webmproject.org/vp9/

either displayed image and dragging the mouse, the other perspective views could be accessed. Additionally, participants could access the refocused images through a slider shown between test and reference, or by double clicking on the point of the image they wished to see in focus. In the second session, participants were first shown an animation of all possible perspective and refocused views. The perspective views were shown in a serpentine order, to mimic natural interaction with parallax effect (see Figure 2). Ten perspective views per second were shown, to ensure a smooth transition. At the end of the perspective views animation, the refocused images were displayed at four frames per second, going from foreground to background and from background to foreground in a smooth transition. The total length of the animation was 13.6 seconds. After the animation was completed, interaction was enabled, allowing participants to change perspective and refocused views as in the first session. No grading was possible before the end of the animation.

Participants were asked to rate the quality of the test stimuli, compared to the uncompressed reference. Before the experiment, a training session was established to acclimatize participants with artefacts and distorsions in the test images. Four training samples, created by compressing one additional content from the dataset on various bitrates, were manually selected by expert viewers, and displayed along with the uncompressed reference.

Randomization was applied on the display order of the stimuli, independently for each session, and the same content was never displayed twice in a row. As each subject took part in both sessions, a break of ten minutes was enforced between the sessions to prevent fatigue.

A laboratory for subjective video quality assessment, set up according to ITU-R Recommendation BT.500-13 [10], was used for the test. In particular, a Samsung SyncMaster2443 monitor of 24 inches and native resolution of 1920 × 1200 pixels was employed. The monitor settings were calibrated 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 controlled lighting system in the room included neon lamps with 6500 K color temperature, while the color of the background walls was mid grey. The illumination level measured on the screens was 15 lux. The distance of the subjects from the screen was approximately equal to 7 times the height of the displayed images, according to requirements in ITU-R Recommendation BT.2022 [11]. Subjects were allowed to move further or closer to the screen.

A total of 21 naive subjects (9 males and 12 females) participated in the test, for a total of 21 scores per stimulus. Subjects were between 18 and 35 years old, with a mean age of 22.3 years at the moment of the test. Before the experiment, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively.

C. Score Processing

Outlier detection and removal was performed on the collected scores, according to ITU-R Recommendation BT.500-13 [10]. One outlier was detected, leading to 20 scores per stimulus. After removing the outlier, the Mean Opinion Score (MOS) and the corresponding 95% Confidence Intervals (CIs) were calculated for each stimulus, separately for each visualization and interaction approach, assuming a Student's t-distribution.

The total time each subject spent on interacting with the perspective and refocused views was computed, following [6], using the formulas:

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

$$\bar{R}_{k,i,j} = \sum_{s=1}^{S} r_{k,s,i,j}, \tag{2}$$

where p and r are the seconds spent on each perspective and refocused view, respectively, $u = 1, 2, \ldots, U$ 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, $j = 1, 2, \ldots, M$ indicates the stimulus, and $k = I, A$ indicates the visualization and interaction approach (I indicating interaction, as used in the first session, and A indicating animation, as used in the second session).

The results were then aggregated to get the total time each subject spent on each stimulus:

$$\bar{T}_{k,i,j} = \bar{P}_{k,i,j} + \bar{R}_{k,i,j}. \tag{3}$$

Finally, the results were averaged across the subjects to obtain their general trend:

$$P_{k,j} = \frac{1}{N}\sum_{i=1}^{N} \bar{P}_{k,i,j}, \tag{4}$$

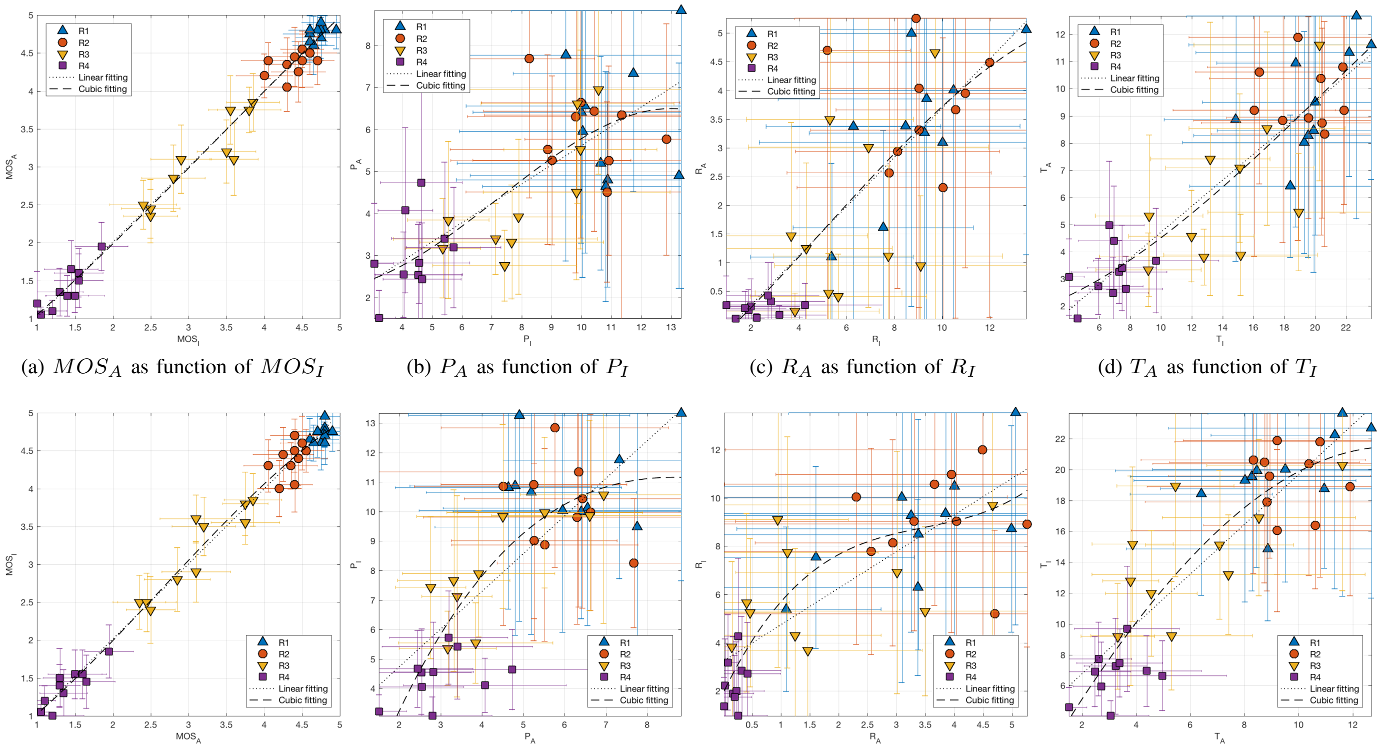
$$R_{k,j} = \frac{1}{N}\sum_{i=1}^{N} \bar{R}_{k,i,j}, \tag{5}$$

$$T_{k,j} = \frac{1}{N}\sum_{i=1}^{N} \bar{T}_{k,i,j}. \tag{6}$$

D. Statistical Analysis

To gain insights on the correlation between the two visualization and interaction approaches, statistical analysis was performed on the subjective scores obtained in the two sessions. In addition, correlation between the number of seconds spent on average on perspective and refocused views, along with the total number of views, was performed for each stimulus ($P_{k,j}$, $R_{k,j}$ and $T_{k,j}$, respectively). For simplicity, from now on we will refer to $P_{k,j}$, $R_{k,j}$ and $T_{k,j}$ as tracking values.

First and third order fittings were applied to the MOS and the tracking values obtained with the two approaches, following the ITU-T Recommendation P.1401[12]. Absolute prediction error (RMSE), Pearson correlation coefficient, Spearman's rank correlation coefficient and Outlier ratio were computed



(a) MOS_A as function of MOS_I (b) P_A as function of P_I (c) R_A as function of R_I (d) T_A as function of T_I

(e) MOS_I as function of MOS_A (f) P_I as function of P_A (g) R_I as function of R_A (h) T_I as function of T_A

Fig. 3: Comparison of MOS and tracking values obtained with different visualization and interaction approaches, along with linear and cubic fittings. Points are differentiated by compression ratio.

for accuracy, linearity, monotonicity and consistency, respectively.

III. Results

Figure 3 shows the scatter plots comparing the MOS and tracking values obtained with the two visualization and interaction approaches. The horizontal and vertical bars represent the CIs corresponding to the values shown in the horizontal and vertical axes, respectively. To improve visualization, points were divided by compression ratio. Additionally, linear and cubic fitting are shown for each comparison. Table II shows the performance indexes, computed on the data pairs $[\widehat{X}_A, X_B]$, in which $\{A, B\} = \{I, A\}$ indicates the two visualization and interaction approach and $X = \{MOS, P, R, T\}$ represent the value under comparison. $\widehat{X}_A$ are the values obtained with approach A after linear and cubic fitting, while X_B are the values obtained with approach B.

As shown in Figure 3 (a) and (e), the MOS values lie on a 45° line, indicating that under identical conditions, the two visualization and interaction approaches tend to give the same value. One-way ANOVA performed on the MOS results grouped by visualization and interaction approach confirms that they are statistically equivalent ($p = 0.9602$). Thus, it can be concluded that choosing one approach over the other is not likely to affect the collected scores. Values of Pearson and Spearman's rank coefficients (> 0.99) further validate the linear correlation between the MOS scores obtained with the two approaches. The results validate the use of interaction-only methodologies as opposed to passive approaches, as having the same visualization experience does not seem to contribute to any change in the scores.

Figures 3 (b) and (f) show the results of the comparison between the tracking values related to the perspective views P_I and P_A. It can be immediately observed that, on average, subjects spent more time interacting with the perspective views in the purely interactive approach, while in the second approach the average interaction time is noticeably reduced. This is easily explained considering that, in the second approach, subjects were already exposed to the collection of perspective and refocused views for almost 14 seconds. Values of Pearson and Spearman's rank coefficients indicate that P_I and P_A are not strongly correlated, as seen in Table II. A slight improvement can be observed in the case of R_I and R_A, which present a stronger correlation (Figures 3 (c) and (g)). However, values of Pearson and Spearman's rank coefficients for T_I and T_A (Figures 3 (d) and (h))indicate that considering the total time of interaction for the whole stimulus improves the correlation (Pearson $= 0.88$, Spearman $= 0.84$ when considering linear fitting).

In general, interaction trends show that, even when presented with an animation of all possible perspective and refocused views, subjects still chose to interact with the content. In particular, confirming a trend already seen in [6], it can be seen that, for both approaches, subjects devoted more time

TABLE II: Performance indexes.

	$[\widehat{MOS}_A, MOS_I]$				$[\widehat{MOS}_I, MOS_A]$			
	Pearson	**Spearman**	**RMSE**	**Outlier Ratio**	**Pearson**	**Spearman**	**RMSE**	**Outlier Ratio**
Linear fitting	0.9922	0.9742	0.1697	5.00%	0.9922	0.9742	0.1696	2.50%
Cubic fitting	0.9923	0.9742	0.1692	5.00%	0.9927	0.9742	0.1643	5.00%
	$[\widehat{P_A}, P_I]$				$[\widehat{P_I}, P_A]$			
	Pearson	**Spearman**	**RMSE**	**Outlier Ratio**	**Pearson**	**Spearman**	**RMSE**	**Outlier Ratio**
Linear fitting	0.7652	0.7066	1.1281	0.00%	0.7652	0.7066	1.8827	7.50%
Cubic fitting	0.7738	0.7017	1.1100	0.00%	0.8065	0.7066	1.7291	7.50%
	$[\widehat{R_A}, R_I]$				$[\widehat{R_I}, R_A]$			
	Pearson	**Spearman**	**RMSE**	**Outlier Ratio**	**Pearson**	**Spearman**	**RMSE**	**Outlier Ratio**
Linear fitting	0.8068	0.7856	1.0469	0.00%	0.8068	0.7856	1.9696	0.00%
Cubic fitting	0.8079	0.7856	1.0441	2.50%	0.8411	0.7856	1.8030	0.00%
	$[\widehat{T_A}, T_I]$				$[\widehat{T_I}, T_A]$			
	Pearson	**Spearman**	**RMSE**	**Outlier Ratio**	**Pearson**	**Spearman**	**RMSE**	**Outlier Ratio**
Linear fitting	0.8805	0.8400	1.5315	2.50%	0.8805	0.8400	2.8113	10.00%
Cubic fitting	0.8823	0.8400	1.5202	2.50%	0.8960	0.8400	2.6327	7.50%

interacting with contents compressed at higher bitrates, while for lower bitrates the time of interaction was limited. For example, refocused views of contents compressed at the lowest bitrate were almost never accessed when using the second approach. The results can be explained considering that, for low bitrates, artefacts are more easily detected; hence, less time is needed for subjects to decide on the score. Moreover, interaction with low quality contents is less likely to be a pleasant experience, which may explain why subjects chose not to engage with that type of content. On the other hand, a more careful inspection may be needed to detect artefacts in contents compressed at higher bitrates. In this case, the animation alone may not be sufficient for subjects to decide which score to assign.

IV. Conclusions

In this paper we described the result of a comparison between two interactive approaches for subjective quality evaluation of light field images. Results show that the subjective scores obtained with the two approaches are statistically equivalent. Although the average time of interaction is sensibly decreased when presenting subjects with an animation prior to the interaction, statistical analysis proves that the tracking values obtained with the two approaches are well correlated. Moreover, it is shown that for both approaches, subjects spent more time interacting with contents compressed with higher bitrates, while for lower bitrates the total time of interaction noticeably decreased.

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).

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] 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.

[3] P. A. Kara, A. Cserkaszky, A. Barst, T. Papp, M. G. Martini, and L. Bokor, “The interdependence of spatial and angular resolution in the quality of experience of light field visualization,” in *3D Immersion (IC3D), 2017 International Conference on*. IEEE, 2017, pp. 1–8.

[4] I. Viola, M. Řeřábek, and T. Ebrahimi, “Comparison and evaluation of light field coding approaches,” *IEEE Journal of selected topics in signal processing*, 2017.

[5] I. Viola, M. Rerabek, and T. Ebrahimi, “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.

[6] I. Viola and T. Ebrahimi, “A new framework for interactive quality assessment with application to light field coding,” in *Applications of Digital Image Processing XL*, vol. 10396. International Society for Optics and Photonics, 2017, p. 103961F.

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

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

[11] 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.

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