# QUALITY ASSESSMENT OF COMPRESSION SOLUTIONS FOR ICIP 2017 GRAND CHALLENGE ON LIGHT FIELD IMAGE CODING

*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

In recent years, the research community has witnessed a growing interest in immersive representations of the real world, such as light field. However, due to the increased volume of data generated in the acquisition, new and efficient compression algorithms are needed to store and deliver light field contents. A Grand Challenge on light field image coding was organised during ICIP 2017 to collect and evaluate new compression algorithms for lenslet-based light field images. This paper reports the results of the objective and subjective evaluation campaign conducted to assess the responses to the grand challenge. An adjectival categorical rating methodology with 7-point grading scale was selected to perform subjective assessments, whereas the objective assessment was conducted using popular image quality metrics. Results show that two proposals have comparable performance and outperform the others across all bitrates.

***Index Terms*—** light field, subjective evaluation, objective evaluation, image coding, image compression.

## 1. INTRODUCTION

Light Field (LF) photography has revolutionized the way scenes are captured and visualized, by storing the direction of light rays along with their intensity. Several methods to acquire LF contents have been proposed in the literature, most notably through the use of multi-camera arrays [1] and handheld plenoptic cameras [2]. As more data is captured when compared to traditional photography, efficient compression algorithms are needed for storage and transmission of LF contents. The ICIP 2017 Grand Challenge on LF image coding, in association with JPEG Pleno Call for Proposals, was issued in January 2017 to collect and evaluate new compression solutions for LF images. The grand challenge was divided into two main tasks, devoted on compressing LF images acquired with two different technologies, namely a plenoptic (lenslet) device and a high-density UHD camera array setup. Due to space constraints, this paper will only focus on the former.

For the lenslet-based challenge, proponents were asked to compress LF images acquired with a Lytro Illum plenoptic camera[1], which uses an array of micro-lenses in front of the main sensor. The data obtained from the sensor, usually referred to as lenslet image, needs to be processed to be properly rendered, via transformation to an explicit 4D LF structure of perspective views [3]. For the challenge, the proponents could follow two workflows: one focused on compressing the lenslet image (Figure 1), and the other focused on compressing the stack of perspective views obtained after transformation to 4D LF structure (Figure 2). Additionally, proponents were asked to provide a renderer, either proprietary or belonging to a third party, that could make the decoded bitstream ready for visualization, supporting their adopted representation model. This step was implemented to collect and assess different representation models for LF rendering.

Overall, a total of five submissions were received as responses for the ICIP 2017 Grand Challenge. Two of the proposals followed the workflow described in Figure 1, whereas three adopted the workflow described in Figure 2. Additionally, two state-of-the-art video codecs were used as anchors to compare and validate the results. Authors of the first algorithm $P01$ exploit the redundacies in the 4D LF structure of perspective views by estimating a part of them as a weighted sum of other perspective views, adopting a linear approximation prior [4]. They use HEVC to encode and transmit part of the views, while non-encoded views are estimated by solving an optimization problem. For algorithm $P02$, authors arrange the perspective views into a multiview structure that can be exploited by the corresponding extension of HEVC, namely MV-HEVC [5]. They also propose a rate allocation scheme to progressively assign the Quantization Parameters (QP) in order to optimize the performance. Authors of $P03$ design a lenslet-based compression scheme that uses depth, disparity and sparse prediction information to reconstruct the final set

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).

[1] https://www.lytro.com/

978-1-5386-1737-3/18/$31.00 

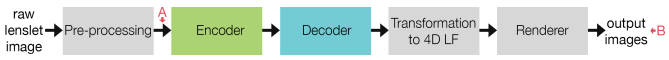



**Fig. 1**: Encoding workflow for lenslet images.

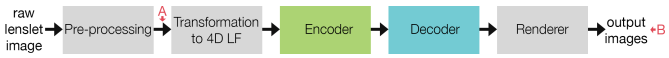



**Fig. 2**: Encoding workflow for perspective views.

of views [6]. The bitrate allocation can be configured to improve the reconstruction by encoding the lenslet image using JPEG 2000, or to allow random access by encoding a subset of views. Authors of $P04$ propose a novel representation of the 4D LF as a multi-modal Gaussian Mixture Model, which can be used to reconstruct the perspective views from the parameters of the model [7]. Their framework can also be employed to produce depth information and apply segmentation. For algorithm $P05$, authors propose a lenslet-based encoding scheme that uses a fully reversible transformation to 4D LF to create sub-aperture views, which are then optimally rearranged and compressed using enhanced illumination compensation in JEM software[2]. Adaptive filtering is then applied to reconstruct the lenslet image [8].

## 2. VISUAL QUALITY ASSESSMENT

All the proposals were assessed through full reference objective metrics and subjective evaluations after the rendering stage (point B in Figures 1 and 2). The reference $B_{Ref}$ was obtained by omitting the encoding and decoding stages in the workflow (shown in green and blue, respectively). Codecs were also evaluated at their maximum reconstruction power $B_{Max}$, obtained similarly by performing an as low as possible compression in the workflow. The evaluation was carried out in three separate steps, to better assess the impact of the compression and the rendering in the final result:

1. $B$ against $B_{Ref}$: Evaluation of the combined impact of encoder, decoder and renderer of the proposed algorithm against the uncompressed rendered content, on four fixed compression ratios.

2. $B$ against $B_{Max}$: Evaluation of the impact of encoder and decoder of the proposed algorithm, using as reference the results of running the encoder at its maximum reconstruction quality $B_{Max}$. This step was implemented to isolate the impact of the proposed renderer on the overall quality.

3. $B_{Max}$ against $B_{Ref}$: Evaluation of the proposed renderer with respect to the reference renderer. This step was implemented to assess the proposed rendering model without the influence of compression artefacts.

[2]https://jvet.hhi.fraunhofer.de/

All three evaluation steps were implemented for the objective assessment, whereas for the subjective assessment the second step was discarded, as changing the reference from $B_{Ref}$ to $B_{Max}$ in the tests would have biased the results.

### 2.1. Dataset and coding conditions

Five contents were selected from an LF image dataset to be compressed for the Grand Challenge, namely I01 = *Bikes*, I02 = *Danger_de_Mort*, I03 = *Stone_Pillars_Outside*, I04 = *Fountain_&_Vincent_2* and I05 = *Friends_1* [9]. The central view of each content is depicted in Figure 3.

Demosaicing and devignetting was applied on the raw camera data to create the 10-bit lenslet images (point A in Figures 1 and 2). Each lenslet image was then processed using the LF MATLAB Toolbox v0.4 [10, 11] to create $15 \times 15$ 10-bit perspective views, which were also color and gamma corrected. Both the lenslet image and the perspective views were given as possible input for the Grand Challenge. The LF MATLAB Toolbox was selected as reference renderer, and the input perspective views constituted the reference $B_{Ref}$. The performance of the proposed coding algorithms was evaluated on four fixed compression ratios, namely $R1 = 0.75$ bpp, $R2 = 0.1$ bpp, $R3 = 0.02$ bpp, and $R4 = 0.005$ bpp. The ratios were computed with respect to the raw lenslet image size ($7728 \times 5368$ pixels).

To assess the performance of the proposals, two anchors were created using state-of-the-art video codecs, namely HEVC Main10 and VP9. Following the workflow depicted in Figure 2, both codecs perform the compression on the perspective views, which were previously rearranged according to a serpentine order, converted to YCbCr format following ITU-R Recommendation BT.709-6 [12], and downsampled from 4:4:4 to 4:2:2, 10-bit depth, little endian format. For the first anchor, the HEVC implementation x265 was used[3], while for the second anchor, the VP9 reference software was used to compress the pseudo-temporal sequence[4]. Full description of the command line used to create the anchors can be found in the JPEG Pleno Lenslet Dataset website[5]. The anchors were not evaluated at their maximum reconstruction power, as the reference renderer was used in the workflow ($B_{Max} = B_{Ref}$). Moreover, due to the limitations of their representation model, the authors of $P04$ chose not to submit any results for compression ratio $R1$. Hence, a total of 160 stimuli were used for the evaluation. A summary of the proposals and the anchors can be found in Table 1.

### 2.2. Objective metrics

To evaluate the impact of the distorsions caused by the proposed algorithms, PSNR and SSIM were selected from the

[3]https://www.videolan.org/developers/x265.html
[4]https://www.webmproject.org/vp9/
[5]http://grebjpeg.epfl.ch/jpeg_pleno/index_lenslet.html

**Table 1**: Summary of compression schemes.

| Proponents | Description |
|---|---|
| **HEVC** | Anchor: Compression of perspective views using HEVC Main10 (x265 software implementation). |
| **VP9** | Anchor: Compression of perspective views using VP9 (reference software). |
| **P01** | Compression of perspective views using HEVC and linear approximation prior [4]. |
| **P02** | Compression of perspective views using MV-HEVC [5]. |
| **P03** | Compression of lenslet image using JPEG 2000 and depth, disparity and sparse prediction [6]. |
| **P04** | Compression of perspective views modeled as Gaussian Mixture Model [7]. |
| **P05** | Compression of lenslet image using optimal arrangement and enhanced illumination model [8]. |

(a) I01 (b) I02 (c) I03 (d) I04 (e) I05

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

literature to objectively assess the visual quality of the contents. The metrics were applied separately to luminance channel $Y$ of each perspective view $(k, l)$, where $k = 1, ..., K$, $l = 1, ..., L$ and $K = L = 15$ represent the total number of perspective views, as generated from the toolbox.

PSNR was computed for chrominance channels $U, V$ of perspective views $(k, l)$, and a weighted average was calculated assigning factor 6 to channel $Y$, and factor 1 to $U$ and $V$ [13]. The average PSNR value for $Y$ channel was then computed across the viewpoint images:

$$\widehat{PSNR}_Y = \frac{1}{(K-2)(L-2)} \sum_{k=2}^{K-1} \sum_{l=2}^{L-1} PSNR_Y(k, l), \quad (1)$$

$\widehat{PSNR}_{YUV}$ and $\widehat{SSIM}_Y$ were analogously computed following Equation 1.

### 2.3. Subjective Methodology

Following the ITU-R Recommendation BT.500-13 [14], a comparison-based adjectival categorical judgement methodology with a 7-point grading scale was selected to perform the subjective visual quality assessment, from -3 (much worse) to +3 (much better), with 0 indicating no preference.

A passive assessment was considered in order to ensure the same experience for all participants [15]. To avoid negative bias in the subjects, only a subset of 97 out of 225 perspective views was presented in the animation, as suggested in [16], since the rest of the views already presents high visual distorsion before compression that can negatively affect the results. As recommended in the aforementioned study, participants were shown the LF contents as pre-recorded animations navigating between the perspective views in a serpentine order, to mimic the parallax effect. The views were displayed at a rate of 10 frames per second (fps), to ensure a smooth transition. The total length of the animation was 9.7 seconds. Each stimulus was displayed alongside the uncompressed reference in a side-by-side arrangement. The position of the reference was fixed for the duration of the test, and participants were informed beforehand on which side of the screen the reference would be displayed.

Participants were asked to rate the quality of the test stimuli when compared to the uncompressed reference. A training session was organized before the experiment to familiarize 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. The experiment was split in four sessions. In each session, the stimuli were shown along with the uncompressed reference, corresponding to approximately 8 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 sessions, hence evaluating all 160 stimuli. A break of ten minutes was enforced between sessions.

The test was conducted in a laboratory for subjective video quality assessment, which was set up according to ITU-R Recommendation BT.500-13 [14]. A professional Eizo ColorEdge CG318-4K 31.1-inch monitor with 10-bit depth and native resolution of $4096 \times 2160$ pixels was used for the tests. The monitor settings were adjusted 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 consisted of adjustable 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 monitor was approximately equal to 7 times the

height of the displayed content, conforming to requirements in ITU-R Recommendation BT.2022 [17]. Subjects were allowed to move further or get closer to the screen.

A total of 28 subjects (19 males and 9 females) participated in the test, for a total of 28 scores per stimulus. Subjects were between 18 and 35, with a mean age of 23.14 years old. Before starting the test, all subjects were examined for visual acuity and color vision using Snellen and Isihara charts, respectively.

### 2.4. Subjective Data Processing and Statistical Analysis

Outlier detection and removal was conducted on the collected scores, according to ITU-R Recommendation BT.500-13 [14]. No outlier was detected, leading to 28 scores per stimulus. The Mean Opinion Score (MOS) was computed for each stimulus, and the corresponding $95\%$ Confidence Intervals (CIs) were calculated assuming a Student's t-distribution.

To determine whether the differences in MOS between the proponents were statistically different, a one-sided Welch's test at $5\%$ significance level was conducted on the results, with the following hypotheses:

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

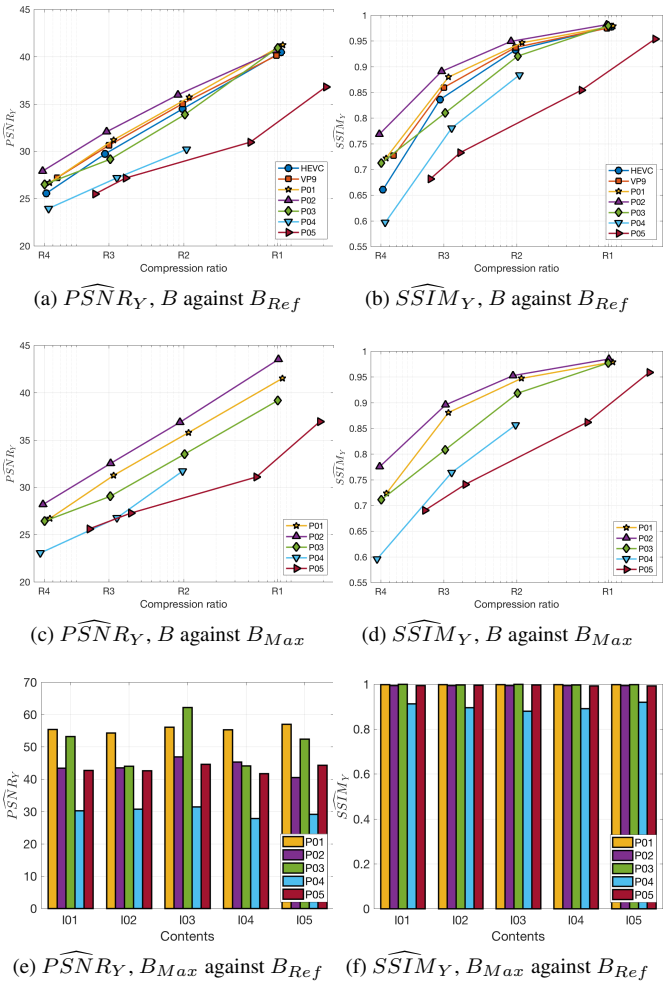
in which $A$ and $B$ are the proposed algorithms under comparison. The test was conducted for each compression ratio and for each content. If the null hypothesis were to be rejected, then it could be concluded that codec $A$ performed better than codec $B$ for the given content and compression ratio, at a $5\%$ significance level. Additionally, a one-way ANOVA test was performed on the results to determine the overall difference between codecs.

## 3. RESULTS

In this section, the results of objective and subjective quality evaluation are outlined. Results of the evaluation campaign are shown in Figures 4, 5 and 6. Results of $\widehat{PSNR}_{YUV}$ are omitted as they exhibited similar trends with respect to $\widehat{PSNR}_Y$.

### 3.1. $B$ against $B_{Ref}$

Results of $\widehat{PSNR}_Y$ and $\widehat{SSIM}_Y$ computed using $B_{Ref}$ as reference (Figure 4 (a) and (b), depicted for content $I02$) show that all codecs have similar performance for compression ratio $R1$, with the exception of $P05$, which is considerably worse. For compression ratios $R2$ and $R3$, codecs $P04$ and $P05$ perform worse than the other codecs, while $P01$ and $P02$ achieve the best results. In particular, $P01$ and VP9 have similar performance, whereas HEVC has a slightly poorer behaviour. For the lowest bitrate, $P02$ clearly outperforms the anchors and other codecs.




(a) $\widehat{PSNR}_Y$, $B$ against $B_{Ref}$ (b) $\widehat{SSIM}_Y$, $B$ against $B_{Ref}$

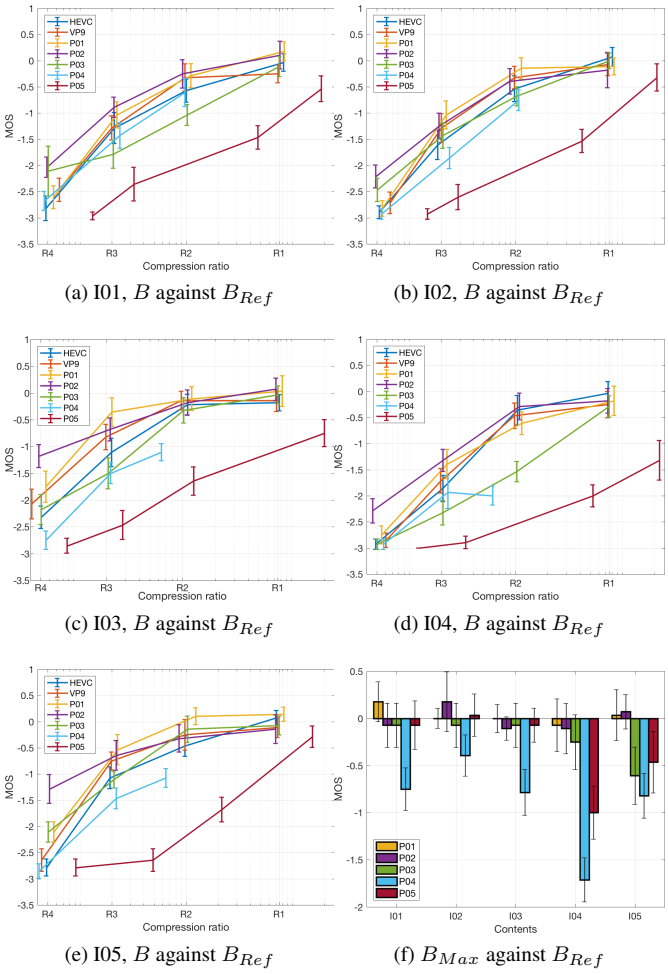
(c) $\widehat{PSNR}_Y$, $B$ against $B_{Max}$ (d) $\widehat{SSIM}_Y$, $B$ against $B_{Max}$

(e) $\widehat{PSNR}_Y$, $B_{Max}$ against $B_{Ref}$ (f) $\widehat{SSIM}_Y$, $B_{Max}$ against $B_{Ref}$

**Fig. 4**: Results of the objective evaluations. The first two rows show metric vs bitrate for representative content I02, the first using $B_{Ref}$ and the second $B_{Max}$ as reference. The third row shows the results of comparing $B_{Max}$ against $B_{Ref}$ for all contents. $\widehat{PSNR}_Y$ and $\widehat{SSIM}_Y$ are used as metric in the first and second columns, respectively.

Results of subjective evaluations confirm the trend. In particular, all codecs have similar performance for the highest bitrate, with the exception of $P05$ (Figure 5 (a - e) and Figure 6 (d)). Among all proponents, $P01$ has the best performance, $P02$ being a close second. For compression ratio $R2$, proponents $P01$ and $P02$ perform similar to anchor VP9 and they surpass the other codecs on more than three out of five contents (Figure 6 (c)). The same trend can be observed for compression ratio $R3$, where $P01$ always performs better than the other codecs, with the exception of $P02$, which has worse results for only one out of five contents (Figure 6 (b)). For the lowest bitrate, $P02$ has the best performance, ranking better than the other codecs on at least three out of five contents, followed by $P03$ and $P01$ (Figure 6 (a)).

One-way ANOVA performed on the results of the subjec-

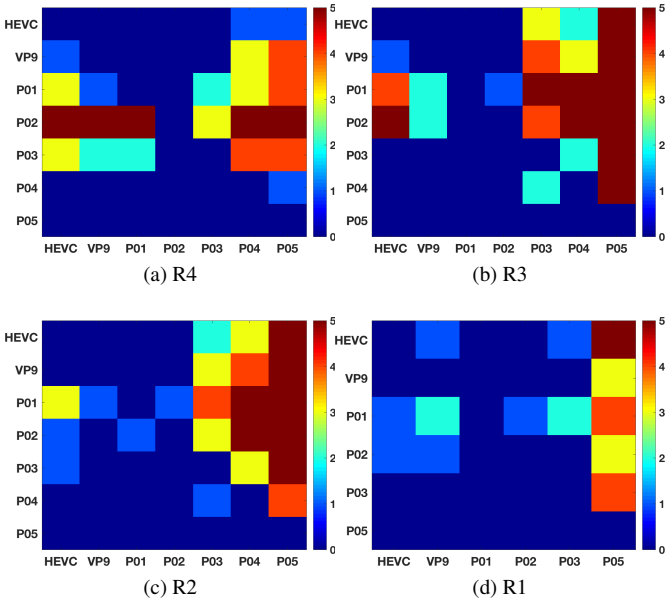



(a) I01, $B$ against $B_{Ref}$ (b) I02, $B$ against $B_{Ref}$

(c) I03, $B$ against $B_{Ref}$ (d) I04, $B$ against $B_{Ref}$

(e) I05, $B$ against $B_{Ref}$ (f) $B_{Max}$ against $B_{Ref}$

**Fig. 5**: Results of the subjective evaluation. MOS vs bitrate for all contents, with respective confidence intervals (a - e), and comparison of $B_{Max}$ with respect to $B_{Ref}$ for all contents (f).




(a) R4 (b) R3

(c) R2 (d) R1

**Fig. 6**: Pairwise comparison results for subjective tests. Each cell represents the number of contents for which $MOS_i$ was found to be statistically better than $MOS_j$; $i$ indicates the row and $j$ the column of the matrix.

tive tests confirms that the codecs are significantly different ($p = 2.1376 \times 10^{-111}$). In particular, proponent $P03$ has comparable performance with respect to the anchors. Proponents $P01$ and $P02$ have statistically equivalent behaviour and they are statistically better than the anchors, whereas $P04$ and $P05$ perform statistically worse than the anchors.

Results show that the chosen encoding workflow does not have a direct influence on the visual quality of the compressed images, as algorithms adopting one or the other workflow can be found among the best and worst performing alike. While state-of-the-art video codecs are crucially employed in the best performing solutions, they result in subpar visual quality in the case of $P05$. This can be explained considering that their algorithm performs the full transformation to 4D LF after compression, which may lead to error propagation.

### 3.2. $B$ against $B_{Max}$

The comparison of the results obtained using $B_{Max}$ as reference (Figure 4 (c) and (d), shown for content $I02$) exhibit similar trends with respect to what has been discussed in Section 3.1, although $P02$ shows a significant gain in performance when using $\widehat{PSNR}_Y$ as metric. It is worth mentioning that, in the $\widehat{PSNR}_Y$ case, proposal $P03$ seems to perform significantly worse when the reference is set to $B_{Max}$ when compared to reference $B_{Ref}$ (Figure 4 (a)), at least for higher bitrates.

### 3.3. $B_{Max}$ against $B_{Ref}$

The objective quality evaluation of $B_{Max}$ against $B_{Ref}$ (Figure 4 (e) and (f)) shows that all proposed renderers achieve favorable results, with the exception of $P04$. However, subjective results show that $B_{Max}$ is never perceived as better than $B_{Ref}$, and in certain cases it is considered as significantly worse than the reference (Figure 5 (f)). In particular, while some proposed renderers were sometimes rated as slighly better than the reference, they fail to be significantly better, as the confidence interval is always seen to be crossing the zero. Moreover, in case of content $I05$, only $P01$ and $P02$ are considered equivalent to the reference, while all other codecs significantly underperform when compared to the reference renderer. Additionally, the renderer proposed in $P04$ is always perceived as worse than the reference. This is mainly due to the fact that the codec uses a mixture of Gaussians to represent the LF structure, leading to poor results when using

full-reference objective metrics.

## 4. CONCLUSIONS

In this paper we report the results of objective and subjective quality assessment of new codecs to compress light field images. Results show that direct application of state-of-the-art video codecs to compress light field images can be improved using new codec designs. In particular, two codecs were found to outperform others in both objective and subjective terms. It was also demonstrated that no proposed representation model is statistically better than that adopted as reference. Finally, it should be noted that, in addition to compression efficiency and visual quality, other criteria such as complexity, delay and random access should be also considered when adopting a preferred solution.

## 5. REFERENCES


[1] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance imaging using large camera arrays," in *ACM Transactions on Graphics (TOG)*. ACM, 2005, vol. 24, pp. 765–776.

[2] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[3] Marc Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, 2006.

[4] Shenyang Zhao and Zhibo Chen, "Light field image coding via linear approximation prior," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

[5] Waqas Ahmad, Roger Olsson, and Mårten Sjöström, "Interpreting plenoptic images as multi-view sequences for improved compression," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

[6] Ioan Tabus, Petri Helin, and Pekka Astola, "Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and JPEG 2000," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

[7] Ruben Verhack, Thomas Sikora, Lieven Lange, Rolf Jongebloed, Glenn Van Wallendael, and Peter Lambert, "Steered mixture-of-experts for light field coding, depth estimation, and processing," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1183–1188.

[8] Chuanmin Jia, Yekang Yang, Xinfeng Zhang, Xiang Zhang, Shiqi Wang, Shanshe Wang, and Siwei Ma, "Optimized inter-view prediction based light field image compression with adaptive reconstruction," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017.

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

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

[11] Donald G. Dansereau, Oscar Pizarro, and Stefan B. Williams, "Linear volumetric focus for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, Feb. 2015.

[12] ITU-R BT.709-6, "Parameter values for the HDTV standards for production and international programme exchange," International Telecommunication Unionn, June 2015.

[13] Jens-Rainer Ohm, Gary J Sullivan, Heiko Schwarz, Thiow Keng Tan, and Thomas Wiegand, "Comparison of the coding efficiency of video coding standards - including high efficiency video coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, 2012.

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

[15] Irene Viola, Martin Řeřábek, and Touradj 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.

[16] Irene Viola, Martin Řeřábek, and Touradj Ebrahimi, "Comparison and evaluation of light field coding approaches," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, 2017.

[17] 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.