# Objective and Subjective Evaluation of Light Field Image Compression Algorithms

Irene Viola*, Martin Řeřábek*, Tim Bruylants†, Peter Schelkens†, Fernando Pereira‡ and Touradj Ebrahimi*
*Multimedia Signal Processing Group (MMSPG), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
†Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Brussels, Belgium
†iMinds, Zwijnaarde, Belgium
‡Instituto Superior Tecnico - Instituto de Telecomunicações (IST/UL-IT), Lisbon, Portugal

***Abstract*—This paper reports results of subjective and objective quality assessments of responses to a grand challenge on light field image compression. The goal of the challenge was to collect and evaluate new compression algorithms for light field images. In total seven proposals were received, out of which five were accepted for further evaluations. For objective evaluations, conventional metrics were used, whereas the double stimulus continuous quality scale method was selected to perform subjective assessments. Results show competitive performance among submitted proposals. However, in low bitrates, one proposal outperforms the others.**

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

## I. Introduction

Light Field (LF) photography aims at expanding the possibilities of traditional photography by capturing information about the direction and the intensity of light rays. This can be achieved by positioning a micro-lens array in front of the image sensor. This way, instead of capturing just the sum of incident light, it is possible to capture the amount of light travelling along the rays composing the scene. As it captures more information about the scene, LF photography creates more data when compared to traditional photography. Therefore, to store and to transmit such images, an efficient compression format is needed.

The ICME 2016 Grand Challenge was issued in January 2016 to collect new compression solutions for LF images, and to evaluate them using both objective and subjective quality assessment methodologies[1]. The grand challenge was focused on compression schemes for raw LF images acquired with a lenslet-based plenoptic camera. More specifically, a Lytro Illum plenoptic camera was used for data acquisition. The participants were requested to compress lenslet images in YCbCr 420 format and 8 bit precision. Lenslet images were created from raw sensor data by applying demosaicing, devignetting, clipping, and color space conversion. The challenge required submission of compression and decompression algorithm capable of processing the given image data according to the end-to-end chain depicted in Fig. 1. Specifically, the proponents were asked to implement steps from A to A'.

The Matlab implementation of the Light Field Toolbox v0.4 [2][3] was exploited to create an LF data structure from the lenslet based LF image. The LF data structure is created by stacking sub-aperture images, each containing those samples from each micro-lens element that are supporting a particular viewpoint. The sub-aperture images obtained can then be rendered on conventional displays. The resulting LF data structure is a 5-D array with dimensions of $15 \times 15 \times 434 \times 625 \times 4$, in which $15 \times 15$ is the number of sub-aperture images, $434 \times 625$ is the resolution of each sub-aperture image, and $4$ corresponds to RGB plus a weighting component.

Overall, seven submissions were received as responses to the call for proposals in the framework of this grand challenge. Only five of them were accepted in the reviewing process for further evaluation. Proponents were assigned a random number (P1 to P5) to anonymize their identity. In general, two main coding approaches were proposed. The first approach uses a modified version of HEVC Intra encoder to compress the lenslet image by exploiting existent redundancies. The second approach creates the LF data structure prior to coding and then rearranges the sub-aperture images in a pseudo-temporal sequence to be coded with HEVC. In the following paragraphs, we present the submitted algorithms in details. The presentation order does not correspond to the label assigned to each codec.

In [4] authors suggested to use HEVC Intra Profile to code the lenslet structure, and to improve its performance by integrating self-similarity compensated prediction and estimation. The proposed solution exploits the correlation between neighbooring micro-images in the lenslet image. The image is partitioned in blocks using HEVC partition patterns. Then, two blocks are selected for predicting the current block, one given by best block matching in the search window and the other selected by searching for best linear combination between the first selected block and a second block in the same window. The best among the two is selected for self-similarity estimation.

The same approach is used in [5], which integrates self-

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), COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET. This work was also partially carried out in the framework of ImmersiaTV under the European Unions Horizon 2020 research and innovation programme (grant agreement no. 688619) and funded by Swiss State Secretariat for Education, Research and Innovation SERI.

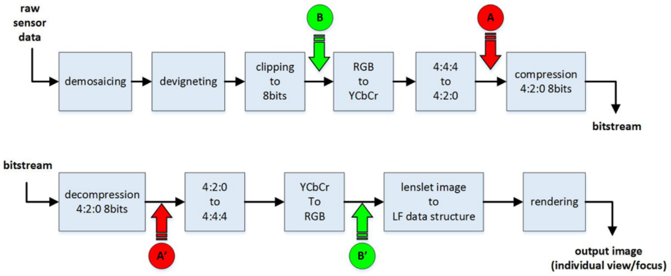


Fig. 1: End-to-end chain for compression and decompression of LF lenslet image.

similarity compensated prediction in HEVC Intra coding, and additionally implements locally linear embedding to further improve the compression performance. Locally linear embedding estimates the current block by solving a least-squares optimization problem to find the best linear combination of $k$ nearest neighbors in a casual search window.

The authors in [6] use HEVC Intra profile to encode the lenslet image; however, the conventional intra prediction from reconstructed information is improved by allowing the predictor to use only blocks from its reconstructed neighbors. In addition to that, advanced motion vector prediction is used.

In [7] the chosen approach is to partition the lenslet image into tiles of equal sizes, which are then ordered in a pseudo-temporal sequence using a properly selected scan order. Then the sequence is encoded using HEVC.

Authors in [8] use a different approach, and propose a compression of LF images based on pseudo-sequences of sub-aperture images. The lenslet image is first converted from YUV420 to RGB444 color space. Then the lenslet is processed to obtain the multiple views that compose the LF data structure. The views are color and gamma corrected and then converted back to YUV420. A subset of them is then rearranged in a specific coding order that accounts for similarities between adjacent views and coded using the JEM encoder[1].

This paper describes and analyses the results of the subjective and objective quality evaluations for LF compression schemes. The proponents were also compared to an anchor generated using legacy JPEG, referred to as P0 in the rest of the paper. The conventional objective metrics were used. The subjective tests were performed using the Double Stimulus Continuous Quality Scale (DSCQS) method [9] and a side-by-side presentation.

## II. Objective evaluation

### A. Dataset and coding conditions

Twelve LF images from a publicly available LF image dataset [10] were selected for the grand challenge. The central view of each content is depicted in Figure 2. The performance of the proposed compression algorithms is evaluated at four fixed compression ratios, namely $R1 = 10:1$ (1 bpp), $R2 = 20:1$ (0.5 bpp), $R3 = 40:1$ (0.25 bpp), $R4 = 100:1$ (0.1 bpp). The full table of contents and coding conditions can be found in [1]. The ratios are computed with respect to the size of the raw data obtained from the camera. To obtain the sub-aperture images suitable to compute the objective metrics, the lenslet images were processed using the LF MATLAB toolbox function *LFDecodeLensletImageSimple* [2][3].

### B. Metrics

To measure distortions introduced by the compression algorithms, the LF data structure, obtained after compressing and decompressing the lenslet image, is compared to the uncompressed reference, which is obtained by omitting the steps from $A$ to $A'$ depicted in Fig.1. The metrics chosen to perform the evaluation are PSNR and SSIM, applied separately to individual color channels. The PSNR is computed on the $Y$ channel as follows:

$$PSNR_Y(k,l) = 10\log_{10}\frac{255^2}{MSE(k,l)}, \qquad (1)$$

in which $k$ and $l$ are the indexes of the sub-aperture images. The $MSE(k,l)$ for each image is computed as follows:

$$MSE(k,l) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j)-R(i,j)]^2, \qquad (2)$$

where $m$ and $n$ are the dimensions of one sub-aperture image (i.e., $n = 625$, $m = 434$). $I(i,j)$ is the $Y$ value for the selected sub-aperture image in the evaluated LF data structure, whereas $R(i,j)$ is the corresponding value in the reference data structure. In the same way, we can compute the PSNR for the other two channels $U$ and $V$, obtained after upsampling the color space as depicted in Fig. 1. A weighted average [11] is then computed as follows:

$$PSNR_{YUV}(k,l) = \frac{6PSNR_Y(k,l)+PSNR_U(k,l)+PSNR_V(k,l)}{8} \qquad (3)$$

The mean of sub-aperture images is subsequentially computed to have an average value for PSNR for $Y$ channel and for $YUV$:

$$PSNR_{X_{mean}} = \frac{1}{(K-2)(L-2)}\sum_{k=2}^{K-1}\sum_{l=2}^{L-1}PSNR_X(k,l), \qquad (4)$$

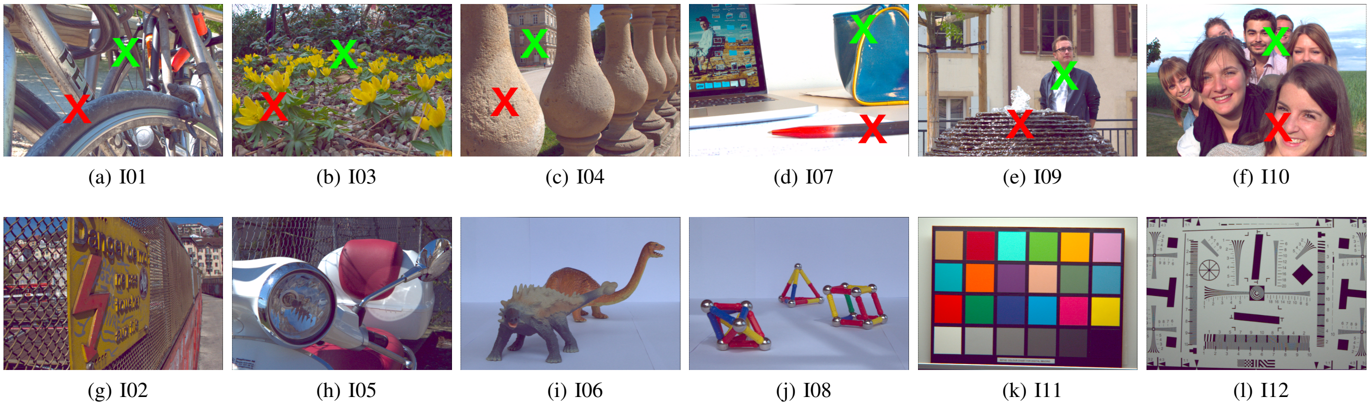
in which $K = 15$ and $L = 15$ represent the number of sub-aperture images, and $X = Y$ and $X = YUV$ for $Y$ channel and for $YUV$ channels, respectively.

In a similar fashion, the SSIM (Structural Similarity Index) is computed on the $Y$ channel of each sub-aperture image as follows:

$$SSIM_Y(k,l) = \frac{(2\mu_I\mu_R+c_1)(2\sigma_{IR}+c_2)}{(\mu_I^2+\mu_R^2+c_1)(\sigma_I^2+\sigma_R^2+c_2)}, \qquad (5)$$

in which $\mu_I$ and $\mu_R$ are the average of the $Y$ channel of the two sup-aperture images at index $k$ and $l$, $\sigma_I^2$ and $\sigma_R^2$ is the variance, and $\sigma_{IR}$ is the covariance of the two sub-aperture images in channel $Y$. $c_1 = (p_1D)^2$ and $c_2 = (p_2D)^2$ are two

[1] https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/tags/HM-16.6-JEM-2.0rc1/


(a) I01 (b) I03 (c) I04 (d) I07 (e) I09 (f) I10

(g) I02 (h) I05 (i) I06 (j) I08 (k) I11 (l) I12

Fig. 2: Central all-in-focus view from each content used in the experiments. Refocused points marked in green (slope 1) and red (slope 2).

variables to stabilize the division; $D$ is the dynamic range of the pixel values, while $p_1 = 0.01$ and $p_2 = 0.03$ by default.

### C. *Analysis and Results*

Figure 3 shows the values of mean PSNR for Y and YUV channels and SSIM for Y channel for content *I03_Flowers*. It can be noticed that for a bitrate of 1 bpp, the difference between the proponents and the anchor is significantly smaller (around 1 dB), whereas it increases for lower bitrates. In general, all the proponents seem to have similar performances for the highest bitrate. P1 performs slightly worse according to PSNR, whereas with SSIM the difference between proposed codecs is negligible. As the number of bit per pixel decreases, P1 outperforms the other codecs, gaining around 3 dB for the same bitrate with respect to P4 in PSNR for Y and YUV channel. Curves for SSIM show similar results, with P1 outperforming all other proponents for lower bitrates.

## III. Subjective evaluation

### A. *Data preparation*

The dataset for the subjective evaluation consists of six LF images, namely, *I01, I03, I04, I07, I09* and *I10*. A thumbnail of the contents is depicted in Figures 2a to 2f. The contents were selected by experts among the twelve contents that were used for objective evaluation.

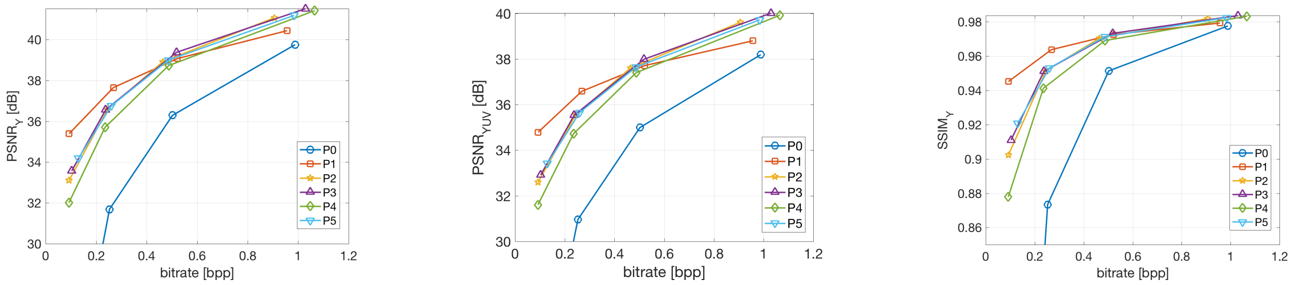
For each content, three all-in-focus sub-aperture images were directly extracted from the LF data structure. From the $15 \times 15$ stack of sub-aperture images, the ones at indexes $(8, i)$, where $i = 5, 8, 11$, were selected to represent different perspectives of each scene. Additionally, the MATLAB toolbox was used to perform a refocus of each scene, using a modified version of the function *LFFiltShiftSum*. This function shifts all the sub-aperture images according to a parameter, referred to as a slope, which determines the focal plane. A sum of the shifted images is performed in order to obtain a single image that is refocused on a specific plane, depending on the value of the slope. The number of images to be shifted and consequently summed defines the Depth of Field (DOF). Summing all $15 \times 15$ images creates the smallest DOF, in which only one specific plane in the image is in focus. On the other side, taking just the central sub-aperture image, which is equivalent to summing just $1 \times 1$ images, brings all the objects in focus (largest DOF). For the test, it has been chosen to sum sub-aperture images from index 5 to index 11 ($7 \times 7$ images) in order to have a larger DOF that still showed the effects of refocusing. Two slopes were selected in order to focus the image on two different planes in the scene. Figures 2a to 2f illustrates the chosen points for refocusing (Slope 1 in green, Slope 2 in red). The values of the slope parameter used in the function are listed in Table I. The three all-in-focus sub-aperture images (perspective views) plus the two refocused images (focus views) form five views per content.

| **Image ID** | **Slope 1** | **Slope 2** |
|---|---|---|
| I01_Bikes | -0.65 | 0.22 |
| I03_Flowers | -0.3 | 0.3 |
| I04_Stone_Pillars_Outside | -0.5 | 0.2 |
| I07_Desktop | -0.5 | 0.5 |
| I09_Fountain_&_Vincent_2 | -0.5 | 0.35 |
| I10_Friends_1 | -0.15 | 0.2 |

TABLE I: Values of slope for refocusing.

### B. *Methodology*

The methodology selected to conduct the subjective tests is based on DSCQS. Two images in native resolution ($625 \times 434$ pixels) were presented simultaneously in a side-by-side fashion. One of the two images was always the uncompressed reference, and its position on the screen was randomized. The other image containing the same perspective or refocus as the reference was compressed by one of the evaluated algorithms at one of the evaluated bitrates. Subjects were asked to rate the quality of both images on a discrete scale from 5 (Excellent) to 1 (Bad). They were informed that one of the images was the reference, but they did not receive any indication whether the reference image was on the left or on the right. Before the experiments, a training session was organized to help subjects to adjust to the change of perspective in the LF structure and the refocusing, and to help them to detect various distortions and compression artifacts. Five training samples from content



(a) $PSNR_Y$ (b) $PSNR_{YUV}$ (c) $SSIM_Y$

Fig. 3: Values of mean PSNR and SSIM for content I03.

*I02* were manually selected by experts. To perform the tests, the QualityCrowd 2 framework [12] was modified to suit the DSCQS methodology.

The experiment was split into four sessions. In each session, 180 pairs of images were shown, corresponding to approximately 45 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 two sessions. A break of ten minutes was enforced between the sessions to avoid fatigue. At the beginning of first session, one dummy example was shown to ease the subject into the task. The resulting scores from dummy stimuli were not included in the final results. Overall, 35 naïve subjects (24 males and 11 females) participated in the subjective experiments, for a total of 17 ratings per LF stimuli. Subjects were between 18 and 33 years old. The average and median age were 22.4 and 22 years old, respectively. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts. Additionally, 18 expert viewers performed the experiments, for a total of 9 additional scores.

### C. Analysis and results

Outlier detection and removal was performed on raw scores of naïve subjects according to the ITU recommendations [9]. One subject was found to be an outlier and the corresponding scores were discarded. This led to 17 scores per stimulus. After outlier removal, the Mean Opinion Score (MOS) was computed for each coding condition $j$ (i.e. for each content, view, proponent and bitrate) as follows:

$$MOS_j = \frac{1}{N}\sum_{i=1}^{N} m_{ij}, \tag{6}$$

where $N$ is the number of subjects and $m_{ij}$ is the score for stimulus $j$ by subject $i$. Figure 4 shows the MOS against bitrate for three of five views evaluated for I03, as well as the average for all views. The proponents and the anchor are plotted with a full line with respective confidence interval, whereas the MOS for the uncompressed reference, with corresponding confidence interval, is shown through a yellow stripe.

In order to determine whether the differences between proponents were statistically significant, all the codecs were compared by means of a two-sided Welch's test at 5% significance level, with following hypotheses:

$$H_0 : MOS_{P_A} = MOS_{P_B}$$

$$H_1 : MOS_{P_A} \neq MOS_{P_B},$$

in which $P_A$ and $P_B$ are the proponents that are being compared. If the hypothesis $H_0$ were to be accepted, it would mean that the difference between means is zero, and that the distribution of difference between mean values follows a t-distribution. On the other hand, if the hypothesis were to be rejected, the conclusion would be that the two values are significantly different. In the test, if the null hypothesis was rejected at $5\%$ significance level, then the two MOS were compared in order to identify which codec performed significantly better. For each content and view, if the hypothesis were to be rejected, the matrix $M$ would be updated as such:

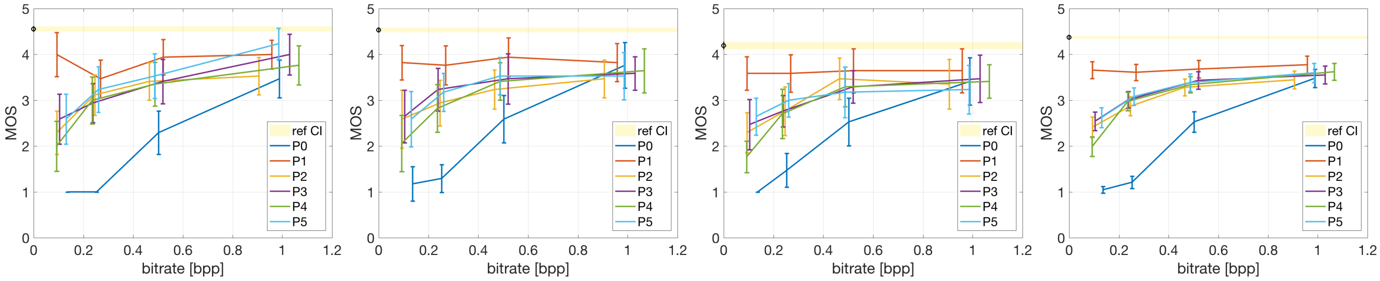
$$M(i,j) = M(i,j) + 1 \text{ if } MOS_i > MOS_j$$
$$M(j,i) = M(j,i) + 1 \text{ if } MOS_i < MOS_j$$

Figure 5 shows for how many contents and views the proponent on the y-axis performs significantly better than the proponent on the x-axis. The minimum value is 0 and the maximum value is 30, corresponding to all possible views and contents.

Similarly to what has been observed in section II-C, all proponents perform similarly for high bitrate and significantly better for low bitrates, when compared to the anchor. For high bitrates, there is no proponent that performs significantly better than the others (Figures 5a and 5b). For lower bitrates, similarly to what has been seen in section II-C, P1 performs better than other proponents, outperforming them for compression rate $R4$ in more than half of the contents (Figure 5d).

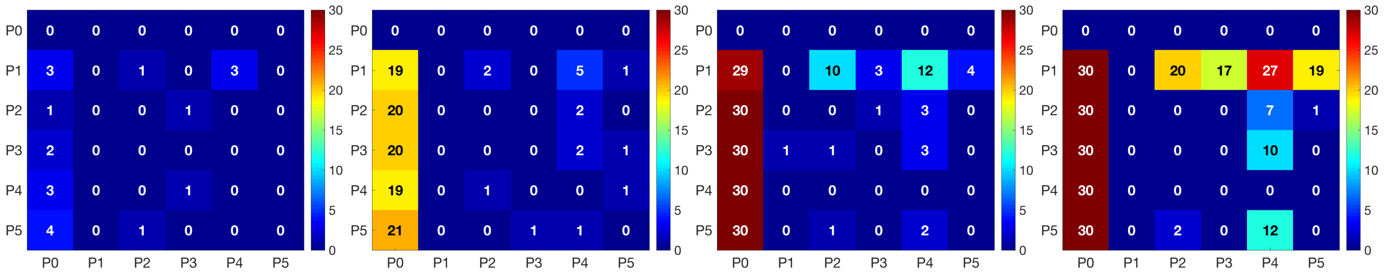
As can be seen in Figure 4c, there is a drop in MOS values for refocused versions of contents, as opposed to all-in-focus views (Figure 4a). The decrease of MOS scores is visible for both, compressed images as well as for uncompressed references. Moreover, the difference of scores between reference and proponents remains constant. These observations suggest that the viewers found that refocusing the content negatively affects its visual image quality.

We performed multiway analysis of variance (ANOVA) on the scores, for different bitrates. The analysis helps determining the difference between means with respect to groups of factors (in this case, contents, proponents and views). We also performed ANOVA on only subgroups of perspective and focus views to better understand the interaction between the groups. The groups "contents" and "proponents" have



(a) Central perspective (b) Side perspective (c) Refocused on background (d) Average

Fig. 4: MOS vs bitrate for content I03.



(a) R1 (b) R2 (c) R3 (d) R4

Fig. 5: Pairwise comparison of codecs for different bitrates.

low p-values associated with them and with their interactions for all bitrates, meaning that at least one of the means for the groups is significantly different from the others. Group "views", however, has interesting results. Although the p-values remain low for almost all bitrates in the group "views", splitting the views in subgroups "perspective" and "focus" leads to remarkable similarities within the subgroups. While different perspectives do not seem to affect the distribution of the scores, a change in focus strongly affects it. This is in agreement with what has been said before about MOS results for refocused views as opposed to all-in-focus views.

Results from expert viewers show similar trends to what has already been said for naïve viewers, although the corresponding confidence intervals are slightly larger, due to the limited amount of scores. For high bitrates, all proponents have equally good performance, and no codec significantly outperforms the others. On the other hand, for low bitrates P1 can be identified as a clear winner, outperforming all the other proponents.

## IV. Conclusion

In this paper we described the results of objective and subjective evaluation of new algorithms for light field image compression, in the framework of ICME 2016 Grand Challenge. Results show that there is much to be gained in using new compression schemes as opposed to legacy JPEG. While for high bitrates all proponents were observed to perform equally well, for low bitrates P1 performed significantly better than all the others.

## References

[1] ISO/IEC JTC 1/SC29/WG1 JPEG, "Grand challenge on light field image compression," Doc. M72022, Geneva, Switzerland, June 2016.

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

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

[4] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–4.

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

[6] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–4.

[7] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–4.

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

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

[10] M. Rerabek and T. Ebrahimi, "New Light Field Image Dataset," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[11] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. 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.

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