Subjective Quality Evaluation of Light Field Data Under Coding Distortions

Emanuele Palma*, Federica Battisti†, Marco Carli†, Pekka Astola* and Ioan Tabus*

* Computing Sciences Unit
Tampere University
Email: firstname.lastname@tuni.fi

† Department of Engineering
Universita degli Studi di Roma TRE
Email: firstname.lastname@uniroma3.it

***Abstract*—This contribution presents the subjective evaluation of the compressed light field datasets obtained with four state-of-the-art codecs: two from the JPEG Pleno Light Field Verification Model and two recent methods for which codecs are publicly available. To the best of our knowledge, currently no subjective testing has been carried out to compare the performances of the four considered codecs. The evaluation methodology is based on Bradley-Terry scores, obtained from pairwise comparisons of the four codecs at four target bit-rates, for four light field datasets. The subset of pairs for which the comparisons are performed is selected according to the square design method, under two design variants, resulting in two datasets of subjective results. The analysis of the collected data, obtained by ranking the subjective scores of the codecs at various bitrates, shows high correlation with the available objective quality metrics.**

I. Introduction

The area of light field or plenoptic image compression is well established, with the last years witnessing an ever increasing number of proposed light field codecs. Dealing with the 4D nature of the recorded images by plenoptic cameras entails many degrees of freedom in designing the image codecs, therefore, the evaluation and comparison of the performance of proposed approaches is not an easy task. The JPEG Pleno light field standardization process has introduced a unifying methodology for reporting objective performance indicators, which are stated in [1].

The subjective evaluation of light field data has been addressed in literature, resulting in few well-accepted testing methodologies, e.g., visually assessing a sequence of angular views of the light field shown as frames in a pseudo-video sequence. In this paper we follow this methodology, in order to assess the quality of recently proposed codecs. Differently than the previous studies, we propose as subjective scoring the Bradley-Terry scores, instead of the MOS scale evaluations which were used in almost all previous papers.

Recently, some works on subjective evaluation of light field and of light field coders have been published. It is useful to underline that, while the methodology for assessing and preserving the quality of image and video is well established [2], for light field data this topic is still under investigation.

In [3] a framework addressing the analysis of the space of attributes for an adequate characterization of light field data is presented. In [4] a dataset for subjective evaluation is presented. In [5], a light field quality dataset is used for understanding the performances of image quality metrics on differently compressed and distorted LF media. A no-reference IQA metric, combining 2D and 3D characteristics of the LF, is introduced too. In [6], the results of subjective and objective quality assessments of compression algorithms for light field images is reported.

In this paper we present the subjective evaluation of coding distortions generated by four codecs at different bitrates. In this work LF images are represented as sub-aperture image (SAI) [1]. The Four light field codecs have been considered: the 4D transform and 4D prediction modes of the JPEG Pleno Verification Model (JPL-VM) 2.1 [7], the light field translation codec (LFTC) [8], and the warping and sparse prediction on regions (WaSPR) [9].

The transform mode of JPL-VM 2.1 directly exploits the 4D structure of the light field using a 4D-DCT. The resulting significant transform coefficients are specified by a hexadeca-tree and entropy coded. In the literature, the 4D transform mode of the JPL-VM 2.1 is also known as the multidimensional light field encoder using 4D transforms and hexadeca-trees (MuLE) [10].

The 4D prediction mode of JPL-VM 2.1 uses a hierarchical disparity-based warping and sparse linear prediction scheme for SAI reconstruction. The disparity and texture data of the lowest hierarchical level, and the prediction error of the texture component in the higher hierarchical levels, are encoded independently with JPEG 2000. The number of SAIs on the lowest hierarchical level depends on the light field characteristics, with dense light fields, such as the EPFL Lytro dataset [11], using only a single reference SAI on the lowest hierarchical level. The 4D prediction mode is also known in the literature as warping and sparse prediction (WaSP) [12].

LFTC uses a sparse set of reference views (encoded with HEVC) in a block-based linear prediction of the full light field, with low-rank approximation and entropy coding of prediction errors. The reference views are divided into blocks that are identified using quad-tree segmentation and SAIs are predicted using a linear combination of the translated blocks. The quad-tree segmentation is used to capture the disparity characteristics of the light field. The prediction errors for each SAI are approximated with principal component analysis and transmitted using entropy coding.

Similar to WaSP, WaSPR uses a hierarchical prediction and coding order, but also improves the coding efficiency with a more refined region-based sparse prediction module followed by inter-coding of prediction errors using HEVC. The region-based sparse prediction module refines the SAI prediction using an additional sparse linear combination of disparity-based regions obtained from neighboring and already decoded SAIs. The inter-coding of the prediction errors improves coding efficiency over the intra-coding scheme of WaSP, especially on dense light fields.

Software implementations of the 4D prediction mode of the JPL-VM 2.1, WaSPR, and LFTC are all publicly available, while MuLE is available as part of the JPL-VM 2.1 software package.

The rest of the paper is organized as follows: Section II gives a description of the proposed methodology. Section III describes the Experimental Setup. Section IV describes the obtained results and Section V draws conclusions.

II. The Proposed Subjective Evaluation Methodology

A. Square-design for reducing the number of pairwise comparisons

In our subjective experiment we denote the adopted codecs as $\mathcal{C}_1, \dots, \mathcal{C}_m$, the tested light field samples as $\mathcal{I}_1, \dots, \mathcal{I}_p$ and for each codec $\mathcal{C}_i$ and each image $\mathcal{I}_j$ we have a set of bitrates $\{r_{ij1}, \dots, r_{ijq}\}$. In general $\{r_{ij1}, \dots, r_{ijq}\}$ are taken to be close to some desired target bitrates, $\{r_1, \dots r_q\}$. For a given image $\mathcal{I}_j$, we identify the set of conditions as $c_1 = (\mathcal{C}_1, r_1), c_2 = (\mathcal{C}_1, r_2), \dots, c_{m \cdot q} = (\mathcal{C}_m, r_q)$. In Figure 1 we show the labeling of the conditions $\{c_1, c_2, \dots, c_{16}\}$ for our case of four codecs and four target rates.

In our study, we also chose to apply pairwise comparison method. This particular method is usually more consistent than standard quality rating schemes. The subjects are prompted to express their binary preference to two compared images. On the results of such method, it then can be applied The Bradley-Terry scoring, which is used to convert user preferences in opinion scores.

The Bradley-Terry scoring is, in fact, used for its convenience, of inferring quality values out of binary quality comparison experiments. However, once the number t of tested conditions grows large, the number of all possible pairwise comparisons, $t(t-1)/2$ grows also very large. It is customary to adopt reduced designs, where not all possible pairs are tested.

For image quality evaluation a traditional reduced design is the square design [13]. This technique was analysed and refined in [14]. Following the conclusions of the study regarding reduced designs in [13], the reduced subset of comparisons can be read from the columns and the rows of a square having the side $\sqrt{t}$, where the conditions are listed in the square in a spiral manner, starting from the leftmost and uppermost corner and spiraling to the center of the square, as we illustrate in Figure 2a, with the 16 conditions of our experiment. The square design requires to perform the binary comparisons of all the conditions in each row, i.e., the six comparisons $(c_1, c_2), (c_1, c_3), (c_1, c_4), (c_2, c_3), (c_2, c_4), (c_3, c_4)$ in row one. Similarly, also the six comparisons of elements are taken for each column. As a result, one has to perform only 48 binary comparisons. We call this subset design as Square design spiral, and use it in Experiment 1.

In this paper we consider also a second type of design, in which the square is filled-in as shown in Figure 2b. After deciding on the elements of the square, the subset of pairs to be tested is formed exactly as in the Square design spiral, with all elements in a row paired with each other, and all elements of a column paired with each other. The reasoning for selecting the filling of the square with conditions as in Figure 2b is that we would like to extract from such experiment valid conclusions for ranking any given codec C_i at all the rates $\{r_1, \dots, r_q\}$, and to see if the quality is increasing with the bitrate, as it is expected. Therefore setting each row for each codec and each column for each target rate one achieves this desideratum. This also ensures a direct comparison of all codecs at each target bitrate r_k. We used this second square design in Experiment$\tilde{2}$.

B. Extracting the Bradley-Terry scores from paired comparison data

The Bradley-Terry scores are very useful statistical quality scores in experiments where there are a number of conditions for which we want to give a score, by performing binary pairwise comparisons of the conditions. If we denote the set of conditions $\{c_1, c_2, \dots, c_t\}$, then for each condition c_i we estimate a score π_i. The Bradley-Terry model assumes that the probability of the binary comparison $X_i > X_j$ of one observation of condition c_i against one observation of condition c_j is given by the model:

$$P(X_i > X_j) = \frac{\pi_i}{\pi_i + \pi_j}.$$

A typical statistical usage of the scores is the evaluation of a tournament where the teams $\{c_1, c_2, \dots, c_t\}$ are playing in pairs several rounds and in the end of the tournament one needs to process the pairwise results to get a ranking of all teams, and even more precisely, a scoring of each team value.

Suppose we observed a number of M pairwise comparisons, and we form and store a matrix W, where the element w_{ij} is the number of times the condition c_i prevailed over condition c_j. Then the maximum likelihood estimates of the quality scores $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots, \pi_t\}$ can be shown to be obtained by maximizing the likelihood function [15]:

$$L(\boldsymbol{\pi}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \left(W_{ij} \log \pi_i - W_{ij} \log(\pi_i + \pi_j) \right). \qquad (1)$$

The Bradley-Terry score for condition c_i is the quantity $\log(\pi_i)$. The maximization of (1) can be achieved using several iterative techniques, see [15] and [16]. For the maximizer of (1) to exist, a technical condition, named "Assumption 1" in [15], is as follows: if one interprets the conditions as nodes in a graph, and draws a directed arc between nodes i and j whenever condition i is preferred over condition j, then we should have a path in the directed graph from any node i to

Codec	WaSP				WaSPR			
Rate	r1	r2	r3	r4	r1	r2	r3	r4
Condition	c1	c2	c3	c4	c5	c6	c7	c8

Codec	MuLE				LFTC			
Rate	r1	r2	r3	r4	r1	r2	r3	r4
Condition	c9	c10	c11	c12	c13	c14	c15	c16

Fig. 1. Definition of the sixteen conditions $c_1, \ldots, c_{16}$, depending on a given codec and a given bitrate

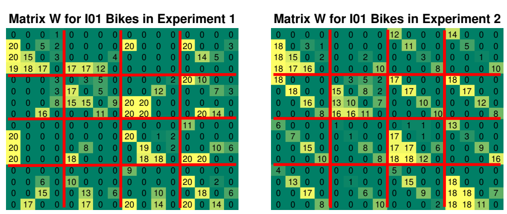
c1	c2	c3	c4
c12	c13	c14	c5
c11	c16	c15	c6
c10	c9	c8	c7

a)

c1	c2	c3	c4
c5	c6	c7	c8
c9	c10	c11	c12
c13	c14	c15	c16

b)

Fig. 2. Square Design: Arrangement of the sixteen conditions in a 4×4 square, for defining the subset of pairs selection. Each condition, situated in cell (i, j) will be paired only to the conditions situated on the row i and on the column j. a) The Spiral Design allocates the conditions ordered increasingly along the red spiral [14]. b) The proposed (Codec,Rate) - Design arranges on each row a codec and on each column a rate, using the conventions from Figure 1.



a) b)

Fig. 3. The collected subjective data is stored in the pairwise comparisons matrix $\mathbf{W}$, where the element W_{ij} is the number of times the condition c_i was preferred to condition c_j. a) Experiment 1, b) Experiment 2.

any node j. Trying to enforce this property will be the goal of the second design in the next subsection.

III. Subjective tests

The designed dataset was used for two subjective experiments: Experiment 1 was conducted in the premises of Roma Tre University and Experiment 2 was conducted at Tampere University. A MatLab Graphical User Interface (GUI) was designed for the experiments. In the following, details of the performed tests are reported.

A. Dataset preparation

The Light Field data considered for the experiment are extracted from the EPFL LYTRO DATASET [11], that is a collection of light field images acquired by using a Lytro camera. In more details: *Bikes*, *Danger de Mort*, *Stone and Pillars Outside* and *Fountain & Vincent 2* have been considered.

For the training session, the *sideboard* light field, extracted from the HCI 4D Light Field Dataset [17], was chosen. The 13x13 central views of each LF dataset are arranged in a snake-like scanning order into a pseudo-video sequence [18], [19]. From the compression side, the four aforementioned codecs were run with four target bitrates, following the [1], but keeping only the bitrates: 0.005, 0.02, 0.1 and 0.75 bpp. We adopted the configuration files provided by the authors of each codec. Overall, we obtained a test set of 16 light fields for training and 64 light fields for test.

With four datasets, collecting a $\mathbf{W}$ matrix with 20 subjects will require 192 comparisons, each pseudo-video sequence lasts 10 seconds resulting in about 80 minutes per subject. A full pairwise experiment instead will require 200 minute, which is exceeding by much the acceptable time for a subject.

B. Subjective experiment description

After an initial screening for normal viewing conditions, 20 volunteers, in an age range between 20 and 61 years old, were selected. The subjects, 15% of women and 85% of men, were considered naive with respect the light field content.

Following the recommendations specified in [20] we set up a test environment with a monitor and a mouse for the user to interact with the test. The background for both the Graphical User Interface (GUI) and the desktop was set to the RGB color [128, 128, 128]. Via the GUI the user could access two sections: Training and Test. In the training section he/she was instructed on the task to be performed by a presentation shown on the screen. After this part, the subject performed the training on the *sideboard* light field image, that was not used for test purposes. The purpose of the training is to make the user get used to the distortions of the compression methods that later he will have to evaluate in the test. After pressing the start button, two light field images in a pseudo-video representation appear showing two differently compressed visual stimuli, since Pair Comparison method relies on comparing two images and expressing a preference. After 10 seconds, the stimuli window closes, allowing the user to cast his/her preference vote. After the training session the user can proceed to the test session. The preferences expressed by the user are recorded in a MatLab matrix. The experiment is divided into two parts to reduce the fatigue of the user. After performing a fixed number of comparisons, the test is over and the user is prompted with the end screen.

IV. Results

The statistical analysis of the collected preference matrix $\mathbf{W}$, was performed by using the approaches proposed in [15], [16].

We checked the Assumption 1 of [15] for the matrices $\mathbf{W}$ in all experiments and all images. Only for the matrix $\mathbf{W}$ shown in Figure 3 a) we found that the assumption is not fulfilled: for the condition c_9 there is a winning path only to c_{13} and for c_{13} there is a winning path only to condition c_9. When calling the MM routine, the report is "no convergence", as expected from [15]. In Experiment 1, the conditions $c_9 = (\mathcal{C}_3, r_1)$ and $c_{13} = (\mathcal{C}_4, r_1)$ were tested against the other codecs only at higher rates, r_2, r_3, r_4, due to the assignment in spiral design, and hence the chance of a subject declaring a win of c_9 or c_{13} over the other conditions was very low, and indeed such win did not happen for the 20 subjects of Experiment 1. On the contrary, for Experiment 2, all the codecs at low bitrates are compared against the other codecs at low bitrates, and indeed each condition was winning over the other tested conditions, at least once (for one subject), making the Assumption 1 true, and no ”no-convergence” results were obtained in Experiment 2, for any of the datasets.

For computing the scores from each matrix W we used three different algorithms from [15] and each time the results were almost identical, except for the image Bikes in Experiment 1, where the algorithm MM [15] did not converge, as expected from the structure of the matrix $\boldsymbol{W}$, explained at end of Section II. Instead we opted to use for this image the results from [16], in which one can still see that the conditions 9 and 13 are outlying values.

We removed from the Experiment 2 two subjects, based on their lower probabilities of concordance with the group majoritarian decisions, lower also than all the probabilities of concordance in Experiment 1. Hence we remained with 20 subjects in Experiment 1 and 18 subjects in Experiment 2.

The obtained Bradley scores are shown in Figures 4 and 5. In Figure 4 there is an almost consistent ranking of the four compression methods at all images and rates, while in Figure 5 the ranking of the compression methods is less consistent.

In Figure 6 we show the results calculated using the SSIM metric for all the images, codecs and bitrates, and we notice the similarity of the ranking of the codecs by the two metrics, especially for Experiment 1, in Figure 4. In the Supplementary Material Document[1] the scores obtained by other two objective metrics, PSNR_Y and PSNR_YUV, are reported.

Figure 7 shows the relation between the subjective score (the Bradley-Terry Score) and the objective metric (SSIM): the cubic trend well describes the overall dependency. The spread of red points (corresponding to Experiment 2) is, in general, narrower than the spread of green points (corresponding to Experiment 1), hence favoring the Experiment 2 design.

Table I shows, in a more quantitative way, the correlation observable in Figure 7.In more details, the correlation between BT Scores and the objective metrics is computed using three correlation coefficients: Pearson Correlation Coefficient (PCC), Kendall Rank Correlation Coefficient (KRCC), and Spearman Rank Correlation Coefficient (SRCC). The correlation between the results of Experiment 1 and 2 is also computed and is seen to be in general stronger than the correlation between the BT scores and any of the objective metrics . Out of the objective metrics, SSIM is the most correlated to the subjective scores for most cases.

V. Conclusions

In this contribution, the subjective evaluation of compressed light field datasets is presented. The experimental results show a high correlation between Bradley-Terry scores and the objective metric for both experiments. The experiments showed slightly different ranking of the compared compression methods, hinting that the reduced design, in which not all

[1]https://www.dropbox.com/s/f91q92m6ijavlk2/Supplementary_Material_Document.pdf?dl=0

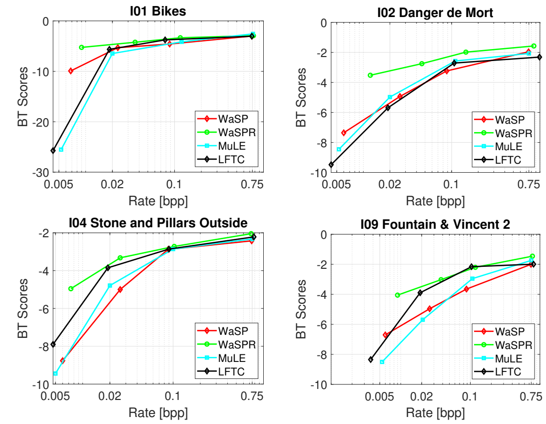


Fig. 4. **Experiment 1** The Bradley-Terry scores obtained for the four images. For each codec the BT score increases with the rate, which shows a desirable consistency of the subjective results. At Bikes, two lowest rates of Mule and LFTC could not be scored reliably, as explained in Section II.

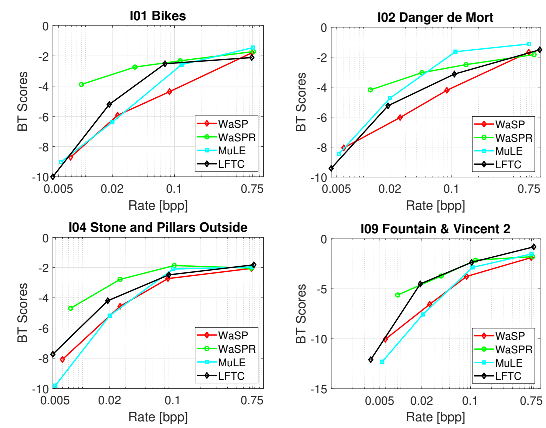


Fig. 5. **Experiment 2** The Bradley-Terry scores obtained for the four images. For each codec the BT score increases with the rate, which shows a desirable consistency of the subjective results.

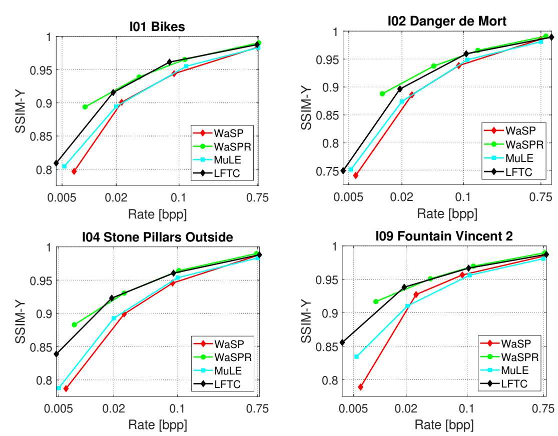


Fig. 6. Objective score SSIM-Y for all four images and all four codecs. The results are correlating well to the subjective scores from Figure 4.

pairs of comparisons are performed, is rather sensitive to the selection of pairs. The obtained reference matrices obtained in Experiment 1 had some problems in the BT-Score conversion, due to its design, but however was showing the most consistent

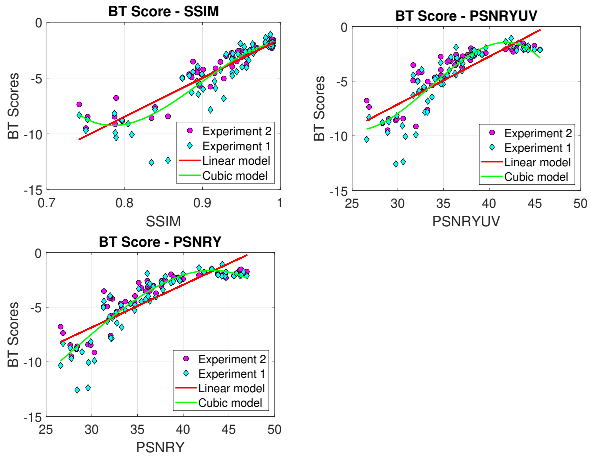


Fig. 7. Bradley-Terry Score against: SSIM (top left), $\mathrm{PSNR_{YUV}}$ (top right), $\mathrm{PSNR_Y}$ (bottom left)

		Experiment 1			Experiment 2		
	Metrics	**PCC**	**KRCC**	**SRCC**	**PCC**	**KRCC**	**SRCC**
I01	$\mathrm{PSNR_Y}$	0.6578	0.7833	0.9382	0.8821	0.7833	0.9294
	$\mathrm{PSNR_{YUV}}$	0.6561	0.8166	0.9529	0.8798	0.8166	0.9441
	SSIM	0.8244	0.8000	0.9441	**0.9647**	0.8000	0.9323
	COR-EXP1/EXP2	**0.8458**	**0.9666**	**0.9941**	0.8458	**0.9666**	**0.9941**
I02	$\mathrm{PSNR_Y}$	0.8536	0.7833	0.9411	0.8850	0.6833	0.8764
	$\mathrm{PSNR_{YUV}}$	0.8529	0.7666	0.9352	0.8971	0.7000	0.8794
	SSIM	0.9631	**0.8166**	**0.9470**	0.9622	0.7166	0.8911
	COR-EXP1/EXP2	**0.9700**	0.7666	0.9029	**0.9700**	**0.7666**	**0.9029**
I04	$\mathrm{PSNR_Y}$	0.8475	0.9166	0.9823	0.8359	0.8500	0.9558
	$\mathrm{PSNR_{YUV}}$	0.8431	**0.9333**	**0.9852**	0.8323	**0.8666**	**0.9617**
	SSIM	0.9842	0.9166	0.9823	0.9734	0.8500	0.9558
	COR-EXP1/EXP2	**0.9892**	0.8666	0.9588	**0.9892**	**0.8666**	0.9588
I09	$\mathrm{PSNR_Y}$	0.8958	0.8500	0.9617	0.8975	0.8500	0.9588
	$\mathrm{PSNR_{YUV}}$	0.8947	0.8666	0.9676	0.8979	0.8666	0.9647
	SSIM	0.9162	0.8666	0.9676	0.9351	0.8666	0.9647
	COR-EXP1/EXP2	**0.9918**	**0.9333**	**0.9852**	**0.9918**	**0.9333**	**0.9852**
Ovr	$\mathrm{PSNR_Y}$	0.6101	0.6984	0.8898	0.8578	**0.7648**	**0.9231**
	$\mathrm{PSNR_{YUV}}$	0.5977	0.6736	0.8725	0.8548	0.7599	0.9227
	SSIM	0.7023	0.7192	0.9036	**0.8925**	0.7341	0.9185
	COR-EXP1/EXP2	**0.7193**	**0.7509**	**0.9175**	0.7193	0.7509	0.9175

TABLE I
CORRELATION OF THE BT SCORES OBTAINED IN EXPERIMENT 1 AND 2 WITH THE OBJECTIVE QUALITY SCORES $PSNR_Y$, $PSNR_{YUV}$, $SSIM$, AND THE CORRELATION VALUES BETWEEN EXPERIMENT 1 AND 2.

ranking of four compression methods. In Experiment 2, we did not have such problems, but the ranking of the compression methods changed from one image to another. The most variations of subjective scores have been observed at low bitrates, where the coding distortions are very noticeable and can be ranked differently by different subjects. For future work, one way to alleviate the drawbacks of reduced designs might be to resort to adaptive selection of the pairs as in [21].

VI. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764951.

REFERENCES

[1] ISO/IEC JTC 1/SC 29/WG 1, "JPEG Pleno Light Field Coding Common Test Conditions," in *ISO/IEC JTC 1/SC29/WG1, Doc. N84027*, July 2019.

[2] F. Battisti, M. Carli, E. Mammi, and A. Neri, "A study on the impact of AL-FEC techniques on TV over IP quality of experience," *Eurasip Journal on Advances in Signal Processing*, 2011.

[3] P. Paudyal, J. Gutiérrez, P. Le Callet, M. Carli, and F. Battisti, "Characterization and selection of light field content for perceptual assessment," in *Int. Conf. on Quality of Multimedia Experience*, 2017.

[4] L. Shan, P. An, D. Liu, and R. Ma, "Subjective evaluation of light field images for quality assessment database," in *Digital TV and Wireless Multimedia Communication*, G. Zhai, J. Zhou, and X. Yang, Eds. Singapore: Springer Singapore, 2018, pp. 267–276.

[5] L. Shan, P. An, C. Meng, X. Huang, C. Yang, and L. Shen, "A no-reference image quality assessment metric by multiple characteristics of light field images," *IEEE Access*, vol. 7, pp. 127 217–127 229, 2019.

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

[7] ISO/IEC JTC 1/SC29/WG1 JPEG, "Verification Model Software Version 2.1 on JPEG Pleno Light Field Coding," in *ISO/IEC JTC 1/SC29/WG1 Doc. N83034*, March 2019.

[8] B. Hériard-Dubreuil, I. Viola, and T. Ebrahimi, "Light field compression using translation-assisted view estimation," *[Proceedings of the Picture Coding Symposium (PCS 2019)]*, p. 5, 2019. [Online]. Available: http://infoscience.epfl.ch/record/270208

[9] P. Astola and I. Tabus, "Coding of light fields using disparity-based sparse prediction," *IEEE Access*, vol. 7, pp. 176 820–176 837, 2019.

[10] M. B. de Carvalho, M. P. Pereira, G. Alves, E. A. B. da Silva, C. L. Pagliari, F. Pereira, and V. Testoni, "A 4D DCT-Based lenslet light field codec," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 435–439.

[11] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Int. Conf. on Quality of Multimedia Experience*,, 2016.

[12] P. Astola and I. Tabus, "Wasp: Hierarchical warping, merging, and sparse prediction for light field image compression," in *European Workshop on Visual Information Processing*, 2018.

[13] O. Dykstra, "Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs," *Biometrics*, vol. 16, no. 2, pp. 176–188, 1960. [Online]. Available: http://www.jstor.org/stable/2527550

[14] J. Li, M. Barkowsky, and P. Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment," in *Proceedings / ICIP ... International Conference on Image Processing*, 09 2012, pp. 629–632.

[15] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *Ann. Statist.*, vol. 32, no. 1, pp. 384–406, 02 2004. [Online]. Available: https://doi.org/10.1214/aos/1079120141

[16] F. Wickelmaier and C. Schmid, "A matlab function to estimate choice model parameters from paired-comparison data," *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, vol. 36, pp. 29–40, 03 2004.

[17] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Computer Vision – ACCV 2016*, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham: Springer International Publishing, 2017, pp. 19–34.

[18] I. Viola, M. Řeřábek, and T. Ebrahimi, "Comparison and evaluation of light field image coding approaches," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1092–1106, Oct 2017.

[19] F. Battisti, M. Carli, and P. L. Callet, "A study on the impact of visualization techniques on light field perception," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 2155–2159.

[20] "Recommendation 500-10: Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500, 2000.

[21] J. Li, M. Barkowsky, and P. Le Callet, "Analysis and improvement of a paired comparison method in the application of 3dtv subjective experiment," in *2012 19th IEEE International Conference on Image Processing*, Sep. 2012, pp. 629–632.