Dictionary Learning-based, Directional and Optimized Prediction for Lenslet Image Coding

Rui Zhong, Ionut Schiopu, Bruno Cornelis, *Member, IEEE*, Shao-Ping Lu, *Member, IEEE*, Junsong Yuan, *Member, IEEE*, Adrian Munteanu, *Member, IEEE*

***Abstract*—In this paper, a novel approach to encode lenslet images is proposed. The method departs from traditional block-based coding structures and employs a hexagonal-shaped pixel cluster, called macro-pixel, as elementary coding unit. A novel prediction mode based on dictionary learning is proposed, whereby macro-pixels are represented by a sparse linear combination of atoms from a generic dictionary. Additionally, an optimized linear prediction mode and a directional prediction mode specifically designed for macro-pixels are proposed. Rate-distortion optimization is utilized to select the best intra prediction mode for each macro-pixel. Experimental results on the EPFL light field image dataset show that the proposed coding system outperforms HEVC and the state-of-the-art in lenslet image coding with an average PSNR gain of 3.33 dB and 1.41 dB, respectively, and with rate savings of 67.13% and 34.30%, respectively.**

EDICS: IMD-CODE image/video coding and transmission

I. Introduction

THE plenoptic camera gained popularity due to its consumer level prices and provided functionalities. In contrast to conventional cameras, which only record light intensity, plenoptic cameras record information about the incoming light from multiple directions, i.e., they provide spatial and angular information in the captured images.

The plenoptic function describes the amount of light travelling through every point in space in any direction at any time instance and over any wavelength [1]. This seven-dimensional function is usually approximated by the Light Field (LF) vector function, using the camera plane and the propagation angles for the primary colors at a given time instance. LF images are captured using one of the following methods: by moving a camera and acquiring images at some specific points in space [2]; using camera arrays [3] to obtain small baseline data known as High Density Camera Array (HDCA) images [4]; using coded apertures [5]; and using microlens arrays, yielding what is known as Lenslet (LL) images [4]. The technological advances in the production of microlens arrays brought by companies such as *Lytro, Inc.* [6], [7] and *Ratrix GmbH* [8] were materialized in consumer-level plenoptic cameras. Such cameras find applications in numerous domains, including image re-focusing [9], image-based rendering [10], computer graphics [11], [12], free-viewpoint video [13], and many more [14].

R. Zhong, I. Schiopu, B. Cornelis, S.-P. Lu, and A. Munteanu are with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium.

J. Yuan is with Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.

Manuscript received August 31, 2017; revised February 26, 2018.

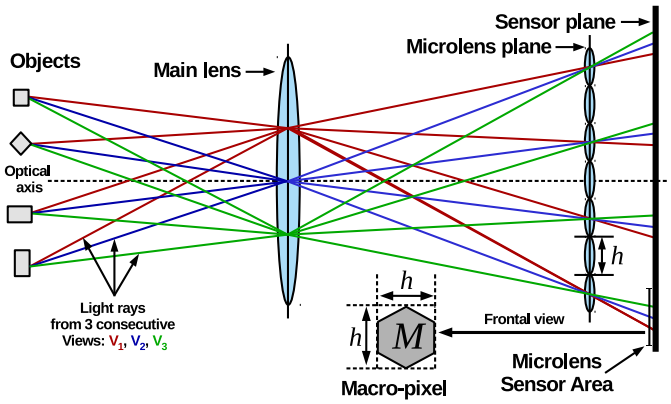

Fig. 1. Illustration of the concept of the plenoptic camera, where the main lens is focusing the light rays reflected by the object onto the microlens plane. The macro-pixel corresponding to a microlens is depicted as a hexagonal-shaped mask of size $h \times h$ which selects the nonzero pixels in the microlens image, where h is the diameter in pixels of the microlens.

Light field cameras can be categorized into **(i)** unfocused plenoptic cameras (e.g. Lytro), introduced by Adelson and Wang [15], and Ng et al. [16], and **(ii)** focused plenoptic cameras (e.g. Raytrix), introduced by Lumsdaine and Georgiev [17] and Perwaß and Wietzke [18]. In this work we address the compression problem for unfocused plenoptic cameras. For these cameras, the main lens is focusing the object's reflected light rays onto the microlens plane, as illustrated in Fig. 1. Each microlens is capturing the converging incoming light rays and is directing them onto the image plane represented by the camera sensor. Each circular microlens of a plenoptic camera produces a so-called macro-pixel [15], [19] which records the incoming light intensity from a discrete set of directions. The overall resolution of the LF image depends on the resolution of each microlens and the microlens array size; for example, the Lytro II camera has a resolution of 40 "Megaray" [20].

Traditional state-of-the-art compression systems were proven to be inefficient when directly applied on lenslet images, due to the inherent spatial discontinuities amongst the macro-pixels. To cope with the large amount of data produced by such cameras, novel compression systems enabling efficient storage and transmission of lenslet images are of paramount importance.

In this paper, a novel compression scheme for lenslet images is proposed. The method introduces a novel prediction

mechanism based on dictionary learning, as well as optimized linear prediction and directional prediction of macro-pixels.

In summary, the novel contributions of this paper are as follows:

1) the use of macro-pixels as elementary coding units [21] instead of traditional block-based coding structures used in conventional codecs such as HEVC;
2) a novel dictionary learning method for macro-pixel prediction;
3) design of optimized linear prediction of macro-pixels; one improves over our previous $L1$ minimization of the prediction error of [21], [22] by accounting for both distortion and rate in the predictor design, not only for the distortion; additionally, both $L1$ and $L2$ distortion metrics are considered, not only $L1$.
4) extension of HEVC's directional intra-modes proposed in our previous work [22] with novel directional macro-pixel prediction modes by employing the concept of multi-hypothesis intra-prediction; to this end, different configurations of neighboring macro-pixels are used as references in directional intra-prediction;
5) optimal rate-distortion selection of the proposed intra-prediction modes and a thorough analysis of the performance provided by these modes;
6) adaptation of HEVC's coding tools to encode residuals for the proposed intra-prediction modes;
7) comparison against state-of-the-art techniques in the literature, demonstrating that the proposed method outperforms the state-of-the-art in lenslet image coding.

The remainder of this paper is organized as follows. Section II discusses the state-of-the-art methods. Section III describes the proposed method. Section IV analyzes the performance of the proposed method. Section V concludes the paper.

II. Related Work

The traditional JPEG standard [23] was tested for lenslet image compression and it proved to be inefficient when applied to this type of images. A powerful alternative is given by the state-of-the-art standard in video coding, namely, High Efficiency Video Coding HEVC [24], which has showed substantially improved compression performance over all its predecessors. However, HEVC was designed with the assumption of local spatial and temporal continuities in video. Since the LF images are characterized by systematic spatial discontinuities between microlens images, the standard HEVC becomes inefficient when encoding this type of data. For an efficient encoding, we find it necessary to adopt the macro-pixel as elementary coding unit and to exploit the inherent redundancies between macro-pixels in the coding system.

Prior art in the area of lenslet image coding includes a variety of techniques. Wavelet compression and intra prediction methodologies were proposed as means for exploiting the intra-frame redundancies. In [25] the authors propose a 4-dimensional Discrete Wavelet Transformation (DWT), combined with the Set Partitioning into Hierarchical Trees (SPIHT) algorithm to code the resulting wavelet subbands. The resulting DWT compression system provides progressive decoding of LF data. An extension is presented in [26] where disparity compensation is performed in the wavelet subbands prior to hierarchical encoding. To further decrease the redundancy within subbands, wavelet compression is applied to viewpoint images generated by extracting corresponding pixels from each microlens instead of the original integral image [27]. In [28], the low-frequency bands, decomposed from reconstructed viewpoint images via a 2D DWT, is coded by a 3D Discrete Cosine Transform (DCT) followed by Huffman coding, while the high-frequency bands are directly processed by arithmetic coding. The above wavelet-based coding schemes provide quality scalability and a complete framework to explore intra-frame redundancies of LF images in the frequency domain.

An alternative to minimize redundancies in the spatial domain is to apply intra prediction directly on microlenses. Self-Similarity (SS) compensated intra-prediction [29] was designed for particular arrangements of microlenses, providing an alternative way to exploit the spatial redundancies in LF images. Bi-directional SS compensation based intra-prediction [30] was proposed to further minimize the prediction error for microlenses with slight view disparities. For the specific rectangle pattern of micro-lenses [29], [30], these SS based intra-prediction methods achieve high coding efficiency and low prediction error. Similarly, a local linear embedding method was proposed in [31] and included into specifically designed HEVC's directional intra prediction modes for rectangle microlenses. Recently, local redundancies were exploited by a Gaussian regression based prediction, integrated into directional intra prediction as a prediction mode [32]. To further explore the repetitive patterns of LF images, in [33] uni-directional and bi-directional SS search based schemes for reference selection compete to decrease the prediction residuals under a Rate-Distortion Optimization (RDO) criterion.

The plenoptic camera can be regarded as an acquisition of conventional 2D images from different viewpoints, at very small distances in between them. In [34], an inter prediction coding method was proposed to capture the redundancies between neighbouring viewpoints. One of the views is intra-coded and serves as reference for predictive coding of the remaining views. Another alternative is to generate multiple viewpoints from a LF image, and to utilize the multiview video coding MVV extension [35] of the HEVC standard on the resulting multiview data. In [36] a 2D warping-based disparity compensation is employed to optimize the prediction, and a linear interpolation is performed to further decrease the coding residual for MVV. To further take advantage of inter-frame and inter-view predictions, a joint motion and disparity estimation method is proposed in [37]. Furthermore, the organization of MVV extracted from LF is regarded as a traveling salesman problem, as well as a lifting transform is applied to obtain disparity compensated LF data, which succeeds to combine the wavelet transform with inter-view prediction [38].

Many contributions were submitted to the ICME 2016 Grand Challenge on LF Image Compression [39]; the bi-predicted SS compensated prediction [29] was one of the accepted contributions. In [40], the plenoptic image is par-

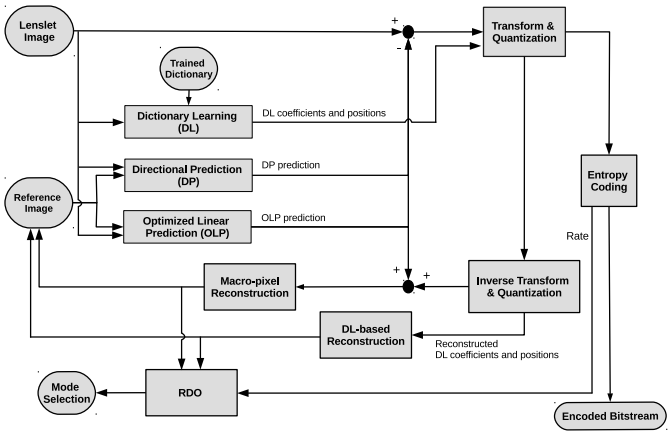

Fig. 2. The proposed coding system whereby dictionary learning-based, directional and optimized linear prediction modes are competing to provide rate-distortion optimized intra prediction for each macro-pixel.

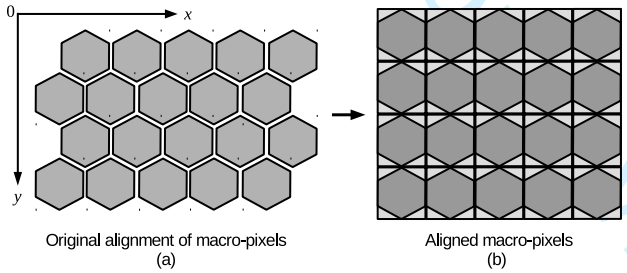

Fig. 3. Example of an 4×5 microlens array. (a) Initial macro-pixel positions. (b) Macro-pixel positions after re-alignment.

titioned into equal tiles, which are scanned in a specific order so that a pseudo video-sequence is generated and used as input for HEVC. In another approach for generating the pseudo video-sequence [41], the conversion from LF image to MVV is carried out by collecting the pixels having the same coordinates in a macro-pixel. In [42], a specific hierarchical reference structure is designed for HEVC-based inter-coding of the pseudo video-sequence.

In our previous work [21], we proposed an L1-optimized prediction algorithm that predicts the macro-pixel as a linear combination of the neighboring reconstructed macro-pixels. This approach exploits the fact that pixels with the same spatial coordinates within neighboring macro-pixels are spatially correlated. In recent work [22], yet to be published, we further reduced the spatial redundancies by designing new HEVC-based directional intra-modes for the macro-pixels. The following section builds on these two approaches bringing substantial improvements over their initial designs. Additionally, a novel intra-prediction method for macro-pixels based on dictionary learning is proposed. Based on these methods, a novel lenslet compression system is devised, where the three types of coding methods are competing in rate-distortion sense. This is detailed next.

III. Proposed Lenslet Compression System

The proposed lenslet image coding system is illustrated in Fig. 2. The system, which follows a closed-loop predictive coding paradigm, takes as input the lenslet image (see Fig. 3(a)) and performs intra-prediction and coding of each macro-pixel. The proposed intra-coding methods, denoted in the following by ξ, include dictionary learning-based intra-prediction, directional prediction, and optimized linear prediction. The coding mode selection is governed by a rate-distortion optimization framework (RDO block in Fig. 2), which provides optimal intra coding for each macro-pixel.

The proposed intra-prediction methods operate in different manners on the macro-pixels. The dictionary learning-based and the optimized linear prediction methods re-align the macro-pixels from the initial lenslet image (see Fig. 3(a)) to a grid structure, as depicted in Fig. 3(b). The directional intra-prediction method is using five other macro-pixel re-alignments, as further detailed in Section III-C.

Compared to our previous design in [21], the optimized linear prediction method proposed in this paper solves a different optimization problem by taking into account the rate needed to encode the residuals and the coding mode. Additionally, the method formulates the optimization problem using both the $L1$ and $L2$ norms, generating two distinct sets of optimized linear prediction modes.

The directional intra-prediction problem, of which incipient results will be published in [22], is further investigated and new neighbourhood configurations are proposed by accounting for the alignment of microlenses in a plenoptic camera.

The paper proposes a novel dictionary learning method, which is employed to learn the basic atoms of a generic dictionary from a training set of LF images, and which is used to linearly represent the macro-pixels based on the learned dictionary. We note that in contrast to the previous two modes, the proposed dictionary-learning based prediction method does not make use of the reconstructed neighbouring macro-pixels to predict the current macro-pixel.

The remainder of this section is organized as follows: Section III-A introduces the proposed dictionary learning-based intra-prediction method; the proposed optimized linear prediction approach is presented in Section III-B; Section III-C describes the proposed directional intra-prediction method; Section III-D presents the adopted entropy coding of intra-coding modes; finally, Section III-E details the rate-distortion-driven selection of the optimal coding modes.

A. Dictionary learning-based method

Dictionary Learning (DL) is a popular methodology that aims at finding a sparse representation of a signal (or collection of signals) by expressing it as a linear combination of only a few atoms from an over-complete dictionary. A crucial part in DL is to define a proper dictionary so that the signal is accurately represented by using the smallest possible number of atoms. A wide range of analytically-defined dictionaries was presented in literature, such as the overcomplete DCT, wavelet and shearlet dictionaries [43], to cite a few. However,

it was shown that learning the dictionary from the signal itself usually yields sparser representations [44].

Given the set of N input signals $\{\mathbf{y}_i\}_{i=1,2,\ldots,N}$, where each signal $\mathbf{y}_i$ contains n data samples, $\mathbf{y}_i = \begin{bmatrix} y(1) & y(2) & \cdots & y(n) \end{bmatrix}^T$, corresponding to a vectorized macro-pixel, the proposed method represents the input matrix $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_N \end{bmatrix}$, of size $n \times N$, using a reduced number of atoms from the dictionary $\mathbf{\Phi}$, of size $n \times d$, where d is the number of atoms in the dictionary. The atoms are selected using the sparse matrix $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix}$, of size $d \times N$, where each sparse vector $\mathbf{x}_i$, of length d, is constrained to have a sparsity s, defined as $\|\mathbf{x}_i\|_0 \leq s$, where $\|\cdot\|_0$ is the ℓ_0 pseudo-norm, so that $\mathbf{x}_i$ is combining only s nonzero elements from $\mathbf{\Phi}$.

The dictionary learning problem can be formulated as:

$$\underset{\mathbf{\Phi},\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{X}\|_F^2 \quad \text{s.t.} \quad \|\mathbf{x}_i\|_0 \leq s \ \ \forall i, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm.

A wide variety of iterative algorithms [45] were proposed in the literature to solve the non-convex problem (1). The usual approach is to alternate between a step that learns the sparse codes and a dictionary update step, like the method of Olshausen and Field [46]. However, to learn the inherent underlying structure of very large datasets, the number of required training samples increases drastically. Consequently, these methods are restrained to work on relatively small patches, extracted from the visual data. Furthermore, traditional dictionary learning methods require updating the entire dictionary repeatedly, which is a costly operation. To alleviate these problems, an Online Sparse Dictionary Learning (OSDL) algorithm is proposed in [47]. OSDL builds structured dictionaries based on the so-called double-sparsity model, which combines a fixed base dictionary $\boldsymbol{\phi}$ with an adaptable sparse component $\mathbf{A}$, i.e., $\mathbf{\Phi} = \boldsymbol{\phi}\mathbf{A}$. The OSDL approach allows for working with larger datasets and it was shown to have a faster convergence rate over traditional dictionary learning methods.

The dictionary learning problem from (1) can be rewritten using OSDL as follows:

$$\underset{\mathbf{A},\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - \boldsymbol{\phi}\mathbf{A}\mathbf{X}\|_F^2 \quad \text{s.t.} \quad \begin{cases} \|\mathbf{a}_i\|_0 = \nu \ \ \forall i \\ \|\mathbf{x}_j\|_0 \leq s \ \ \forall j \end{cases}, \tag{2}$$

where ν is the sparsity for $\mathbf{A}$, and the base dictionary $\boldsymbol{\phi}$ consists of cropped fully separable wavelets, enabling a multiscale analysis, free of border artefacts.

In this paper, a novel intra-prediction mode is proposed by combining RDO with online sparse dictionary learning.

Specifically, the input signal $\mathbf{Y}$ is reconstructed as $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{\mathbf{y}}_1 & \hat{\mathbf{y}}_2 & \cdots & \hat{\mathbf{y}}_N \end{bmatrix} = \boldsymbol{\phi}\mathbf{A}\mathbf{X}$, where only the nonzero elements found in $\mathbf{X}$ are transmitted to the decoder using their positions in $\mathbf{X}$ and their values. Therefore, the following information needs to be transmitted to the decoder:

(i) the nonzero coefficient values stored in the coefficient matrix, denoted by $\mathbf{X}_{nonz}$, which are transformed, quantized, and encoded using CABAC [24] operating with various values of the quantization parameter QP, denoted here for simplicity by q;

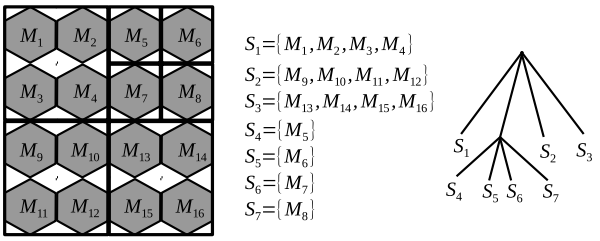

Fig. 4. The partition of $N = 16$ macro-pixel samples into an optimized quadtree structure with $N_m = 7$ nodes.

(ii) the position of the nonzero coefficients, which are marked in the nonzero label positions matrix, denoted by $\mathbf{P}_{nonz}$, and which are losslessly encoded to guarantee an accurate reconstruction of the input signal.

The dictionary learning problem is now formulated as:

$$\underset{\mathbf{A},\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - \boldsymbol{\phi}\mathbf{A}\mathbf{X}\|_F^2 + \lambda R_{DL} \text{ s.t. } \begin{cases} \|\mathbf{a}_i\|_0 = \nu \ \forall i \\ \|\mathbf{x}_j\|_0 \leq s \ \forall j \end{cases}, \tag{3}$$

where the signal is reconstructed at a distortion level, D_{DL}, given by:

$$D_{DL} = MSE(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{i=1}^{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_F^2. \tag{4}$$

The rate R_{DL} necessary to attain the distortion D_{DL} is given by:

$$R_{DL} = R_P(\mathbf{P}_{nonz}) + R_X(\mathbf{X}_{nonz}), \tag{5}$$

where $R_P(\mathbf{P}_{nonz})$ and $R_X(\mathbf{X}_{nonz})$ are the number of bits needed to encode $\mathbf{P}_{nonz}$ and $\mathbf{X}_{nonz}$, respectively. The two rate components are determined as:

$$R_P(\mathbf{P}_{nonz}) = \alpha_1 \cdot N_m, \tag{6}$$

$$R_X(\mathbf{X}_{nonz}) = \alpha_2 \cdot q + \beta_2, \tag{7}$$

where α_1 is a parameter which depends on the sparsity level s and the length of the coefficient vectors d, computed as $\alpha_1 = s \log_2 d$; (α_2, β_2) is the pair of parameters of the least-squares regression line [48] used to encode the coefficients; N_m is the number of vectors from $\mathbf{X}_{nonz}$ transmitted to the decoder.

Our tests have shown that the similar characteristics amongst neighbouring macro-pixels yield redundancies in $\mathbf{P}_{nonz}$. Here a Quadtree merging algorithm [48] is introduced to reduce the redundancy of $\mathbf{P}_{nonz}$, by merging neighbouring position vectors. The idea is to make a binary decision for merging four neighbouring macro-pixels at each node by minimizing a Lagrangian cost function. The optimized tree structure is established with N_m branches. The procedure associates to each branch a list $S_i, i = 1, 2, \ldots, N_m$, where S_i contains the list of macro-pixels grouped by the quadtree such that they are associated to one position vector. Fig. 4 shows an example of merging applied to a set of $N = 16$ macro-pixels.

A representative selection of lenslet images is used as input for the dictionary learning training procedure, which computes a generic dictionary Φ used for all images in the test set. Since the dictionary is common for all test images, it does

not need to be transmitted. For each encoded lenslet image, a specific coefficient matrix $\mathbf{X}$ is computed and encoded using the presented algorithm, so that the decoder can reconstruct the macro-pixels via linear combinations of the selected atoms.

The dictionary learning-based method yields one intra-prediction mode, which is collected by ξ and characterized by the pair (R_{DL}, D_{DL}).

B. Optimized linear prediction method

The second intra-prediction method proposed in this paper is based on the concept of adapting the coding architecture to the lenslet image characteristics by employing the macro-pixels as basic prediction unit. Here, the currently predicted macro-pixel, denoted by $\mathbf{T}$, is predicted linearly based on the three closest macro-pixels in its causal neighborhood, denoted by $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$, which correspond to the Northern, Western and Northwestern macro-pixels. In this paper, a generic macro-pixel with the label T has a corresponding vector $\mathbf{T}$ of length n, which collects the values of the pixels contained in the macro-pixel mask. Therefore, the reconstructed macro-pixel, $\tilde{\mathbf{T}}$, is expressed as a linear combination of $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$, of the form $\tilde{\mathbf{T}} = \mathbf{M}\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is the weight vector defined as $\boldsymbol{\omega} = \begin{bmatrix}\omega_1 & \omega_2 & \omega_3\end{bmatrix}^T$, $\{\omega_1, \omega_2, \omega_3\} \in \mathbb{R}$, $\sum_{i=1}^{3} \omega_i = 1$, and $\mathbf{M} \in \mathbb{R}^{n\times 3}$ is a matrix which collects the causal neighbourhood of T as $\mathbf{M} = \begin{bmatrix}\mathbf{M}_1 & \mathbf{M}_2 & \mathbf{M}_3\end{bmatrix}^T$.

The optimization problem consists of searching for the optimal weight vector, $\boldsymbol{\omega}^*$, for which the residual $\|\mathbf{T} - \tilde{\mathbf{T}}\|_p$ is minimized, where p denotes the norm used in the optimization problem. The Optimized Prediction (OP) mode introduces in the competition, both the $L1$ norm, by setting $p = 1$, and the $L2$ norm, by setting $p = 2$. Hence, the optimization problem is formulated as:

$$\min_{\boldsymbol{\omega}} \|\mathbf{T} - \mathbf{M}\boldsymbol{\omega}\|_p^2 \quad \text{s.t.} \quad \sum_{i=1}^{3} \omega_i = 1. \tag{8}$$

Encoding the original weights $\boldsymbol{\omega}^*$ obtained by solving (8) for each macro-pixel would require a large rate. Therefore, the weights $\boldsymbol{\omega}^*$ are subsequently clustered by means of the K-means clustering algorithm [49] for an efficient encoding. Let B denote the total number of clusters. The indexing of the weights is done according to each of the B cluster centers. During the intra prediction process of each macro-pixel, the B prediction modes are traversed and the best linear prediction mode is determined as:

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}\in\boldsymbol{\mathcal{W}}}{\operatorname{argmin}} \|\mathbf{T} - \mathbf{M}\boldsymbol{\omega}\|_p^2 + \lambda R_{OP}, \tag{9}$$

where R_{OP} is the rate of encoding in the OP mode, $\boldsymbol{\mathcal{W}} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \cdots, \boldsymbol{\omega}_B\}$ is the set of trained prediction weights (i.e. the cluster centers resulting from K-means clustering), and $\hat{\boldsymbol{\omega}} = \boldsymbol{\omega}_o$ is the optimal weight vector with the position index $o \in \{1, 2, \ldots, B\}$ in $\boldsymbol{\mathcal{W}}$.

R_{OP} depends on both the cost of coding the residual, $\mathbf{T} - \tilde{\mathbf{T}} = \mathbf{T} - \mathbf{M}\boldsymbol{\omega}$, and on the cost of encoding the index of the weight vector $\boldsymbol{\omega}$ in $\boldsymbol{\mathcal{W}}$; formally, R_{OP} is thus given by:

$$R_{OP} = R_{res}(\mathbf{T} - \tilde{\mathbf{T}}) + R_{ind}(\boldsymbol{\omega}), \tag{10}$$

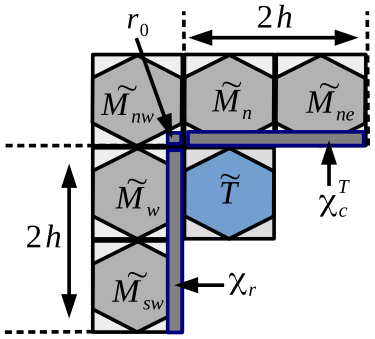

Fig. 5. Sample selection of the $\boldsymbol{\chi}_c$ and $\boldsymbol{\chi}_r$ vectors. Note that $\tilde{M}_{sw}$ is unavailable at the decoder, and it is replaced with a copy of one of the macro-pixels from the rows above it.

where $R_{res}(\mathbf{T} - \tilde{\mathbf{T}})$ is the rate of encoding the residual $\mathbf{T} - \tilde{\mathbf{T}}$ using CABAC, and $R_{ind}(\boldsymbol{\omega})$ is the rate of encoding the index positions of the weight vector $\boldsymbol{\omega}$ in $\boldsymbol{\mathcal{W}}$.

The proposed optimized linear prediction method yields two sets of intra prediction modes, which are collected by ξ. More exactly, a set of B modes is obtained by using the $L1$ norm in (8), and another set of B modes is obtained using the $L2$ norm; in our experiments, we set $B = 32$.

The goals of our design are to either sparsify the prediction error (when using the $L1$ norm) or to minimize the mean square error of the prediction error (when using the $L2$ norm). One cannot say beforehand which norm yields the best performance, as rate also matters. To this end, the RDO (rate-distortion optimization) module decides the best mode in rate-distortion sense.

C. Directional prediction method

The third proposed intra-prediction method, dubbed Directional Prediction (DP), eliminates the spatial redundancies in macro-pixels by estimating the samples in the target macro-pixel based on the previously reconstructed neighbouring macro-pixels. The inherent spatial discontinuities between neighbouring macro-pixels in LF images break the assumption of local spatial continuity employed in conventional block-based codecs, such as HEVC. That is, directly applying the directional intra-prediction modes of HEVC proves to be inefficient on LF images. Hence, new directional prediction methods for LF images are needed that aim at capturing the spatial redundancies by means of directional prediction.

The proposed DP method can be summarized in three steps: **(i)** macro-pixel extrapolation; **(ii)** sample selection; **(iii)** directional prediction.

Macro-pixel extrapolation

In the proposed DP method, in the first step, the reconstructed macro-pixels are extrapolated to generate corresponding $h\times h$ blocks ($h = 17$ for the Lytro camera). Let us consider the generic case of a reconstructed macro-pixel, denoted as $\mathbf{M}$, and let $\tilde{\mathbf{M}}$ denote the extrapolated version of it. The macro-pixel extrapolation process contains a vertical extrapolation followed by a horizontal extrapolation, whereby the pixels in blank areas are copied from the closest boundary pixels.

A horizontal extrapolation for the current pixel position (x, y), at row y and column x, is performed as follows:

$$\tilde{M}(x, y) = M(x_z, y), \tag{11}$$

where $M(x_z, y)$ is the closest available pixel in $\mathbf{M}$, found on column x_z and on the same row y. Note that due to the symmetrical shape of the macro-pixel, $x_z > x$, for $x < \frac{h}{2}$, and $x_z < x$, for $x > \frac{h}{2}$.

The vertical extrapolation is done in a similar way, this time on column x, where $M(x, y_z)$ is the closest available pixel in $\mathbf{M}$, found on row y_z and on the same column x.

The resulting blocks $\tilde{\mathbf{M}}$ are aligned in a grid structure (see Fig. 5 for an example of a possible alignment). In principle, the current macro-pixel, denoted by $\tilde{\mathbf{T}}$ (see Fig. 5), is reconstructed by using the neighboring macro-pixels $\tilde{\mathbf{M}}_n, \tilde{\mathbf{M}}_w, \tilde{\mathbf{M}}_{nw}, \tilde{\mathbf{M}}_{ne}, \tilde{\mathbf{M}}_{sw}$, found at the Northern (n), Western (w), Northwestern (nw), Northeastern (ne), and Southwestern (sw) positions respectively from the current macro-pixel $\tilde{\mathbf{T}}$. We note that, due to the image row-wise scan, $\tilde{\mathbf{M}}_{sw}$ is unavailable at the decoder, therefore it is replaced with a copy of one of the above macro-pixels, i.e, in this case $\tilde{\mathbf{M}}_w$ or $\tilde{\mathbf{M}}_{nw}$.

Sample selection

The main idea of the method is to generate two sets of samples, one on the horizontal direction and one on the vertical direction, used by the directional prediction method (see Fig. 5). The sample found on the bottom-right corner of $\tilde{\mathbf{M}}_{nw}$, denoted $r_0 = \tilde{\mathbf{M}}_{nw}(h, h)$, is also used by the method. The samples on the horizontal direction are collected in a vector $\boldsymbol{\chi}_c$ of length $2h$, while the samples on the vertical direction are collected in a vector $\boldsymbol{\chi}_r$ also of length $2h$. Fig. 5 shows how the two vectors are set using the denoted neighborhood. The vector $\boldsymbol{\chi}_c$ is set using the last row, $x = h$, from the reconstructed macro-pixels $\tilde{\mathbf{M}}_n$ and $\tilde{\mathbf{M}}_{ne}$, and it is defined as

$$\boldsymbol{\chi}_c = \big[M_n(1,h) \;\; M_n(2,h) \;\; \cdots \;\; M_n(h,h) \;\; M_{ne}(1,h) \;\; M_{ne}(2,h) \;\; \cdots \;\; M_{ne}(h,h)\big]^T. \tag{12}$$

Similarly, the vector $\boldsymbol{\chi}_r$ is set using the last column, $y = h$, from the reconstructed macro-pixels $\tilde{\mathbf{M}}_w$ and $\tilde{\mathbf{M}}_{sw}$, and it is defined as

$$\boldsymbol{\chi}_r = \big[M_w(h,1) \;\; M_w(h,2) \;\; \cdots \;\; M_w(h,h) \;\; M_{sw}(h,1) \;\; M_{sw}(h,2) \;\; \cdots \;\; M_{sw}(h,h)\big]^T. \tag{13}$$

In this paper, we introduce five Neighborhood Configurations (NC) for the current macro-pixel. Let us consider the original alignment shown on the left of Fig. 6, where the current macro-pixel, $\mathbf{T}$, has a causal neighborhood of five macro-pixels: $\{\mathbf{M}_i\}_{i=1,2,\ldots,5}$, four of them placed on the previous row and one place on the current row, to the left of $\mathbf{T}$. After applying extrapolation we obtain the set $\{\tilde{\mathbf{M}}_i\}_{i=1,2,\ldots,5}$ and $\tilde{\mathbf{T}}$. Fig. 6 shows the proposed five NC, denoted NC1, NC2, …, NC5, which are obtained as follows:

- *Shift the previous row to the right* with $\frac{h}{2}$ and set:
  - $r_0 = \tilde{\mathbf{M}}_1(h, h)$;
  - for generating $\boldsymbol{\chi}_r$: $\tilde{\mathbf{M}}_n = \tilde{\mathbf{M}}_2$ and $\tilde{\mathbf{M}}_{ne} = \tilde{\mathbf{M}}_3$;
  - for generating $\boldsymbol{\chi}_c$:

  (NC1) $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_5$ and $\tilde{\mathbf{M}}_{sw} = \tilde{\mathbf{M}}_5$;
  (NC2) $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_1$ and $\tilde{\mathbf{M}}_{sw} = \tilde{\mathbf{M}}_5$;
- *Shift the previous row to the left* with $\frac{h}{2}$ and set:
  - $r_0 = \tilde{\mathbf{M}}_2(h, h)$;
  - for generating $\boldsymbol{\chi}_r$: $\tilde{\mathbf{M}}_n = \tilde{\mathbf{M}}_3$ and $\tilde{\mathbf{M}}_{ne} = \tilde{\mathbf{M}}_4$;
  - for generating $\boldsymbol{\chi}_c$:

  (NC3) $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_5$ and $\tilde{\mathbf{M}}_{sw} = \tilde{\mathbf{M}}_5$;
  (NC4) $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_2$ and $\tilde{\mathbf{M}}_{sw} = \tilde{\mathbf{M}}_5$;
- *Keep the original alignment* and set:
  - $r_0 = \tilde{\mathbf{M}}_2(h, h)$;
  - for generating $\boldsymbol{\chi}_r$: $\tilde{\mathbf{M}}_n = \big[\tilde{\mathbf{M}}_2(\frac{h}{2} : h, :) \quad \tilde{\mathbf{M}}_3(1 : \frac{h}{2}, :)\big]$ and $\tilde{\mathbf{M}}_{ne} = \big[\tilde{\mathbf{M}}_3(\frac{h}{2} : h, :) \quad \tilde{\mathbf{M}}_4(1 : \frac{h}{2}, :)\big]$;
  - for generating $\boldsymbol{\chi}_c$:

  (NC5) $\tilde{\mathbf{M}}_w = \big[\tilde{\mathbf{M}}_2(:, \frac{h}{2} : h) \quad \tilde{\mathbf{M}}_5(:, 1 : \frac{h}{2})\big]$ and $\tilde{\mathbf{M}}_{sw} = \big[\tilde{\mathbf{M}}_5(:, \frac{h}{2} : h) \quad \tilde{\mathbf{M}}_5(:, 1 : \frac{h}{2})\big]$.

Directional prediction

The directional prediction of a sample found at row y, and column x in $\tilde{\mathbf{T}}$ is computed as follows:

$$\tilde{\mathbf{T}}(x, y) = \big[\mathbf{P}_k^T \boldsymbol{\chi}_c + \mathbf{Q}_k^T \boldsymbol{\chi}_r\big], \tag{14}$$

where the vectors $\mathbf{P}_k$ and $\mathbf{Q}_k$, of length $2h$, are the k^{th} parameter sets, $k = x + y \cdot h$, with two nonzero elements, and $[\cdot]$ is the rounding function. The pair $(\mathbf{P}_k, \mathbf{Q}_k)$ depends on the pixel's coordinates, $0 \le x, y \le h$, and the angle $g_i = \frac{e_i}{32}$ associated with the directional mode with index i, where e_i is the angular number corresponding to g_i, as shown in Fig. 7. The two nonzero elements are located at the consecutive positions n and $n+1$, and are denoted by d_n and d_{n+1}. Their values are computed as follows:

$$n = x + \lfloor y \cdot g_i \rfloor, \tag{15}$$

$$d_n = \begin{cases} y \cdot g_i + \big\lceil |y \cdot g_i| \big\rceil, & g_i < 0 \\ y \cdot g_i - \lfloor y \cdot g_i \rfloor, & g_i > 0 \end{cases}, \tag{16}$$

$$d_{n+1} = 1 - d_n. \tag{17}$$

d_n is employed to determine which vector contains the two nonzero elements:

$$d_n \in \mathbf{Q}_k, \forall g_i \in \big[H - 27 \quad H + 32\big], i = 2, 3, \ldots, 18, \tag{18}$$

$$d_n \in \mathbf{P}_k, \forall g_i \in \big[V - 32 \quad V + 32\big], i = 19, 20, \ldots, 36. \tag{19}$$

For each of the five NC, the DP method generates h intra-prediction modes using $\boldsymbol{\chi}_l$, h intra-prediction modes using $\boldsymbol{\chi}_r$, and one using r_0, resulting in a total of $2 \cdot h + 1 = 35$ intra-prediction modes for each NC. Moreover, the traditional DC and planar modes from HEVC are also included in competition together with the new DP modes proposed above.

This results in $5 \cdot (2 \cdot h + 1) + 2 = 177$ intra-prediction modes for the DP method, which are all collected by ξ. The index of directional prediction, i, is set: $i = 0$ for the HEVC's DC mode, $i = 1$ for the HEVC's planar mode, and $i = 2, 3, \ldots, 36$ for the DP modes described by (18) and (19). If the new DP modes are selected, the NC is encoded using separate symbols (see Section III-D).

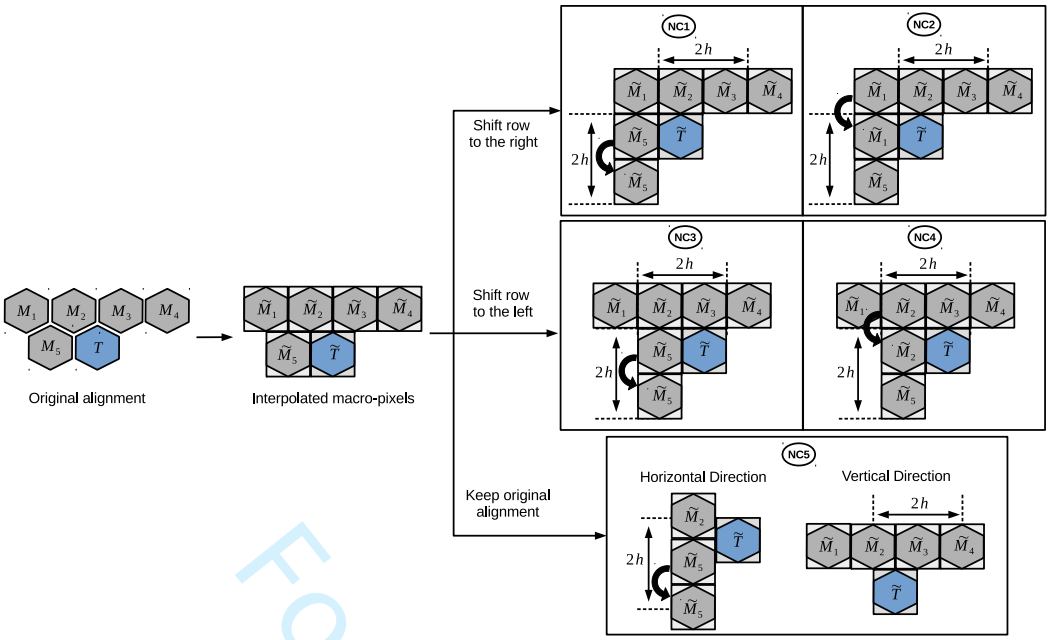

Fig. 6. The 5 cases of neighbourhood configurations for the directional intra-prediction mode. The original alignment found in the plenoptic camera is shifted with $\frac{h}{2}$ to the left, to the right, or keep the same positions. For the cases which involved shifting, the western ($\tilde{\mathbf{M}}_5$) and northwestern ($\tilde{\mathbf{M}}_1$ or $\tilde{\mathbf{M}}_2$) macro-pixels are used to generate two different configurations.

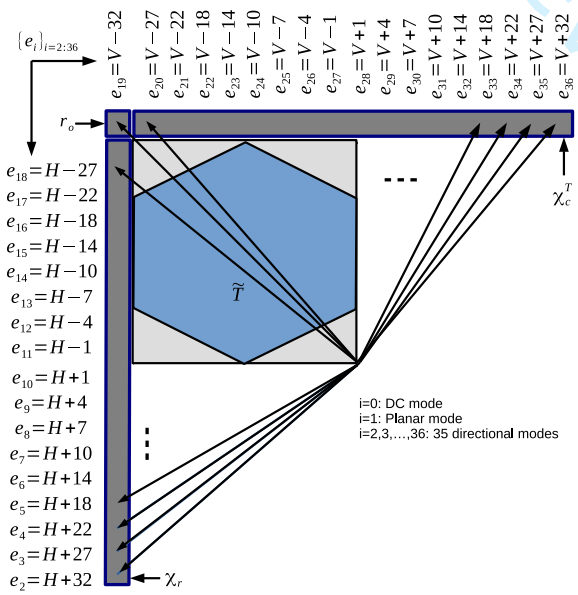

Fig. 7. Directional prediction modes for encoding the current macro-pixel $\mathbf{T}$, corresponding to the directional prediction with indexes $i = 2, 3, \ldots, 36$.

One may note that the proposed Directional Prediction (DP) method has a similar design to the intra prediction method in classical codecs. However, the key concept of the proposed DP method is to derive directional predictors by employing various sets of reference pixels, i.e., by employing multi-hypothesis intra-prediction. The DP mode aims at capturing directional features in macro-pixels. It does succeed in doing so, but it may not necessarily survive the RDO competition against the other methods.

D. Entropy coding of intra-prediction modes

The encoder processes blocks of $N = 16$ macro-pixels at a time. For the DL method, each macro-pixel within the block is independently predicted, as detailed in Section III-A. The OP and DP prediction methods employ decoded macro-pixels located in the causal neighbourhood of the macro-pixel being predicted, as detailed in Sections III-B and III-C respectively.

The HEVC standard was modified and we added the necessary syntax elements to implement the proposed codec. The original syntax elements of HEVC intra prediction consist of the CU_skip_flag, the Most Probable Modes, and the block residual coefficients for intra prediction [50], [51], and are now encoded here using CABAC.

The index of the intra prediction mode is encoded using the syntax index mpm_idx using an alphabet of symbols $\{0, 1, \ldots, 69\}$, where one symbol corresponds to the DL-based mode, 32 symbols to the OP mode, and 37 to the DP mode. If an OP mode is selected, a binary vector, denoted π_{OP}, is collecting an index selection between $L1$ and $L2$ norm. If a DP mode is selected, a NC vector, denoted π_{DP}, is collecting an index selection for the corresponding DP neighborhood configuration. Both π_{OP} and π_{DP} are encoded using an Adaptive Markov Model (AMM) [52] of order 2, with an alphabet of 2 and 5 symbols, respectively.

The HEVC closed-loop coding paradigm, i.e., entropy decoding followed by inverse quantization and transformation, are included in the encoder and are performed to generate the reconstructed macro-pixels. The encoding and decoding processes are part of the loop; this guarantees that correct encoding is performed by matching the encoder with the corresponding decoder, even if the processing unit is changed

TABLE I
ICME 2016 IMAGE TEST IMAGES

Image ID	Image name
$I01$	Bikes
$I02$	Danger_de_Mort
$I03$	Flowers
$I04$	Stone_Pillars_Outside
$I05$	Vespa
$I06$	Ankylosaurus_&_Diplodocus_1
$I07$	Desktop
$I08$	Magnets_1
$I09$	Fountain_&_Vincent_2
$I10$	Friends_1
$I11$	Color_Chart_1
$I12$	ISO_Chart_12

from a block to a macro-pixel.

E. Rate-distortion optimization

The optimal intra-prediction mode is selected as the mode which yields optimal rate-distortion performance. This is determined by solving the following minimization problem:

$$k^* = \underset{k \in \{DL, OP, DP\}}{\operatorname{argmin}} (D_k + \lambda R_k), \quad (20)$$

where D_k and R_k are the distortion and rate respectively associated with mode k selected among the Dictionary Learning (DL) mode (see Section III-A, eqs. (3)-(7)), the $2 \cdot B$ Optimized Linear Prediction (OP) modes (see Section III-B, eqs. (8)-(10)), and the 177 Directional Prediction (DP) modes (see Section III-C, eqs. (11)-(19)) respectively. We note that the rate needed to entropy code the mode itself (see Section III-D) is also accounted for in the above minimization problem.

IV. EXPERIMENTAL EVALUATION

A. Experimental Setup

The same evaluation procedure as in [56] is followed in the experimental evaluation of the proposed coding system, i.e., we use the EPFL test set consisting of 12 LF images to evaluate the coding performance. Table I presents the most used associated labeling for the images in the EPFL test set, where each raw image has a resolution of 7728×5368 pixels, which require $51,854,880$ bytes for storage. The raw images are first demosaiced, devignetted, clipped from 10-bit to 8-bit representation, color calibrated, and converted to the YCBCR4:2:0 color-map representation.

Images 'House and Lake', 'Palais du Luxembourg', 'Red and White Building', and 'Sophie and Vincent 1' from the dataset presented in [53] were used to train the dictionary. The resulting dictionary was used for all the other test images in the dataset. To assess the robustness of the proposed DL-based prediction method against changes in the training set, we have carried out experiments using 5 different training sets, each set containing 4 images selected from the 12 images in the EPFL dataset, the other 8 images being used for testing. We have trained 5 different dictionaries and computed the performance differences between the average PSNR obtained with the proposed method and the PSNR obtained with the reference methods. The experiments reveal that, for each dictionary, the PSNR differences relative to the reference methods are maintained. It is also important to point out that the performance differences relative to the reference methods have a very small variance: the PSNR does not change with more than 0.2 dB relative to the mean.

To demonstrate the advantages of the proposed compression method, we compare the PSNR and the codelength of the encoded LF images against the following coding systems: **(i)** HEVC operating in intra-mode [57], serving as reference codec for the consumer market, denoted here HMINTRA; **(ii)** LLE and SS compensated prediction [54], denoted here SSPRED; **(iii)** the state-of-the-art pseudo-sequence-based compression of [41], denoted here PSEUDOSEQ; **(iv)** the pseudo-sequence-based 2D hierarchical coding structure of [42], denoted here PSEUDOH; **(v)** the recent work described in [55], denoted here as TIP2018; **(vi)** our previous method presented in [22], denoted L1PRED. The experiments are performed using 4 QPs, namely $22, 27, 32$, and 37. Compared to HMINTRA, we notice substantial performance improvements achieved by the proposed method; an attempt to ameliorate HEVC's performance on these images included using CU sizes of 16×16 pixels obtained by zero-padding the 13×13 macro-pixels used in the PSNR evaluations. The results remained modest for HEVC, indicating that such conventional codec designs are not sufficient for lenslet image coding

Let us denote $PSNR_c$ the PSNR of the color channel c of the decoded LF image. $PSNR_c$ is computed relative to the raw 8-bit image as $PSNR_c = 10 \cdot \log_2 \frac{255^2}{MSE}$, where

$$MSE = \frac{1}{Q} \sum_{p=1}^{n_M} \left(\mathbf{M}_p - \tilde{\mathbf{M}}_p \right)^2, \quad (21)$$

where $Q = n_M \times m$ is the total number of pixels in the LF image; m is the number of pixels selected by the macro-pixel (for the Lytro camera $m = 199$); $n_M = 434 \times 541$ is the number of microlenses in the camera configuration, i.e., a microlens matrix of 434 rows and 541 columns obtained after the macro-pixel alignment procedure; $\mathbf{M}_p$ is the original macro-pixel and $\tilde{\mathbf{M}}_p$ is the reconstructed macro-pixel. The reported PSNR is calculated on YUV channels for the 13×13 macro-pixels, for the proposed as well as for all the reference techniques, following the recommendations in [39]. In the RDO block, the Lagrange multiplier λ depends on the selected QP values and is computed as originally proposed in [58]:

$$\lambda = 0.85 \cdot 2^{\frac{QP-12}{3}}. \quad (22)$$

The generic dictionary used in the proposed method, consisting of $d = 256$ atoms, was trained on 4 images selected from a different dataset [53], and used for the compression of the 12 lenslet images in the EPFL test set [56]. Since the resulting generic dictionary is known both by the encoder and decoder, the entire bit-length for the DL-based method accounts only for the cost of encoding the sparse coefficient matrix. One notes that increasing the size of the dictionary increases the rate needed to encode the locations of the non-zero coefficients. We found experimentally that a dictionary of $d = 256$ atoms was sufficient to provide a good rate-distortion trade-off and competitive performance

TABLE II
PERFORMANCE GAINS OF THE PROPOSED METHOD COMPARED TO OTHER METHODS, IMAGES 'HOUSE AND LAKE', 'PALAIS DU LUXEMBOURG', 'RED AND WHITE BUILDING', AND 'SOPHIE AND VINCENT 1' FROM [53] ARE USED TO TRAIN THE DICTIONARY

Img. ID	vs. HMIntra		vs. SSpred [54]		vs. PseudoSeq [41]		vs. PseudoH [42]		vs. TIP2018 [55]		vs. L1Pred [22]	
	PSNR gain (dB)	Bitrate saving (%)	PSNR gain (dB)	Bitrate saving (%)	PSNR gain (dB)	Bitrate saving (%)	PSNR gain (dB)	Bitrate saving (%)	PSNR gain (dB)	Bitrate aving (%)	PSNR gain (dB)	Bitrate saving (%)
$I01$	3.61	−62.98	2.21	−50.42	0.79	−23.93	0.70	−19.02	0.12	−31.44	1.07	−26.83
$I02$	2.63	−53.72	1.88	−43.48	1.61	−42.44	1.29	−38.51	1.18	−34.10	1.37	−34.03
$I03$	2.27	−46.61	1.88	−41.92	1.39	−36.36	1.34	−35.91	0.97	−26.58	1.34	−31.46
$I04$	2.22	−45.26	1.94	−42.14	0.89	−19.34	0.87	−17.96	0.19	8.15	0.98	−22.74
$I05$	3.18	−71.38	2.18	−60.51	1.95	−57.48	1.71	−55.54	1.42	−48.66	0.83	−28.16
$I06$	3.86	−85.39	2.04	−73.92	0.68	−30.74	0.49	−26.09	−0.63	38.24	0.26	−9.45
$I07$	2.41	−54.97	2.30	−46.50	3.04	−65.64	2.62	−63.46	3.39	−72.29	0.57	−17.50
$I08$	3.76	−90.61	3.11	−83.71	2.80	−78.60	2.49	−78.55	2.68	−80.23	0.52	−30.03
$I09$	5.20	−72.86	2.01	−53.63	1.29	−41.17	1.26	−39.96	−1.15	42.49	0.82	−22.87
$I10$	2.70	−66.28	2.58	−64.47	2.78	−70.19	2.65	−69.12	2.75	−69.80	1.31	−37.80
$I11$	4.15	−81.30	1.32	−47.96	4.27	−77.56	4.24	−77.15	4.08	−77.61	0.95	−31.26
$I12$	3.93	−74.17	1.58	−48.82	1.65	−53.27	1.64	−52.95	1.89	−59.79	1.13	−30.93
Avg.	3.33	−67.13	2.09	−54.79	1.93	−49.73	1.78	−47.85	1.41	−34.30	0.93	−26.92

of DL-prediction against the other intra-coding modes. In this paper, for the DL-based method, we used $(\alpha_2, \beta_2) = (-0.042, 1.839)$ and the sparsity $s = 8$, i.e., the positions and the values of only 8 coefficients are transmitted to the decoder for each macro-pixel at each specific QP, while the reference algorithms transmit the residues for m pixels. The 8 coefficient values are encoded relative to their mean, i.e., for all the macro-pixels encoded using the DL-based method, the mean values are collected in a matrix and are transformed, quantized, and encoded using CABAC. Since the coding precision of the mean has a great impact on the rate-distortion performance, the mean values are encoded at a QP value, denoted here $\bar{q}_{QP}$, which is different than the QP value used for the rest of the image. In our experiments, we used $\begin{bmatrix}\bar{q}_{22} & \bar{q}_{27} & \bar{q}_{32} & \bar{q}_{37}\end{bmatrix} = \begin{bmatrix}0 & 2 & 6 & 12\end{bmatrix}$, corresponding to the four test QP values of $22, 27, 32$, and 37 respectively.

B. Experimental results and analysis

Table II reports the BD-PSNR and BD-RATE computed using Bjontegaard's evaluation tools [59] and Fig. 8 shows the RD curves, for the 12 LF images from the EPFL dataset. One can notice from the figures that the proposed method has a better compression performance than that of reference algorithms. Overall, the average PSNR gain is 3.33 dB, 2.09 dB, 1.93 dB, 1.78 dB, and 1.41 dB, against HMINTRA, SSPRED [54], PSEUDOSEQ [41], PSEUDOH [42], TIP2018 [55], respectively, corresponding to 67.13%, 54.79%, 49.73%, 47.85%, and 34.30% in terms of rate savings.

Compared to our previous L1PRED codec [22], the proposed coding system yields an average PSNR gain of 0.93 dB and rate savings of 26.92%. The proposed coding system introduces a dictionary learning-based method, replaces the $L1$ minimization of the residual exploited in [22] by solving completely different optimization problems that employ both distortion and rate and make use of the $L1$ and $L2$ norms, and proposes five different neighbourhood configurations for directional prediction. Fig. 8 shows that at medium and high rates, the proposed method reaches much better compression performance compared to our previous design in L1PRED. In particular, notable PSNR gains at higher bitrates demonstrate the advantages of mode competition brought by the proposed codec.

Fig. 9 shows the color-coded comparison and Fig. 10 presents the results of the comparison between the three proposed intra-coding methods and the HEVC-based intra prediction method, for the EPFL dataset and for the four QP selected values. The results show that the DL-based mode is selected between 20% to 55% for small QP's which is decreasing to 5% to 20% for large QP's; HEVC-based intra prediction methods are selected between 15% and 55% of the cases; the DP method is selected between 15% and 40% of the cases, while the OP method has a selection rate ranging between 10% to 30% of the cases.

One can also notice that: *(i)* the DL-based mode is very competitive for small QP's, while for large QP's the rate of encoding the positions and the values of the nonzero coefficients becomes too large compared to the other modes; *(ii)* the DP and OP mode selection is increasing almost linearly when increasing the QP values, replacing the DL-based mode; *(iii)* as seen in Fig. 9, the DP and OP modes are mostly used inside objects and are replacing the DL-based mode at large QP's; HEVC's DC and planar modes are used around sharp edges, while the DL-based mode is used in flat areas, e.g., see the $I04$ image in Fig. 9 (third row, second column).

Fig. 11 shows the selection of each NC for the DP method. One can notice that NC1 is the configuration selected most of the time, i.e., between 35% and 45% of the cases, NC3 is the second most-selected configuration, between 25% and 30%, followed by NC2 and NC4, each between 10% and 20%, while NC5 is selected between 5% to 10% of the cases. These results are due to the configuration of the 5 neighbourhoods, where for NC1 and NC3 we used $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_5$, while for the other cases, one of the two macro-pixel from the previous row and closest to $\tilde{\mathbf{M}}_5$, is shifted to the $\tilde{\mathbf{M}}_w$ position, i.e., $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_1$ or $\tilde{\mathbf{M}}_w = \tilde{\mathbf{M}}_5$, see the original alignment in Fig. 7. Although NC5 is the closest configuration to the microlens camera configuration, it is the least selected configuration. This is the result of merging two neighboring half macro-pixel blocks resulted after an horizontal or vertical splitting of blocks (see Fig. 6). Hence, the results in Fig. 11 show that

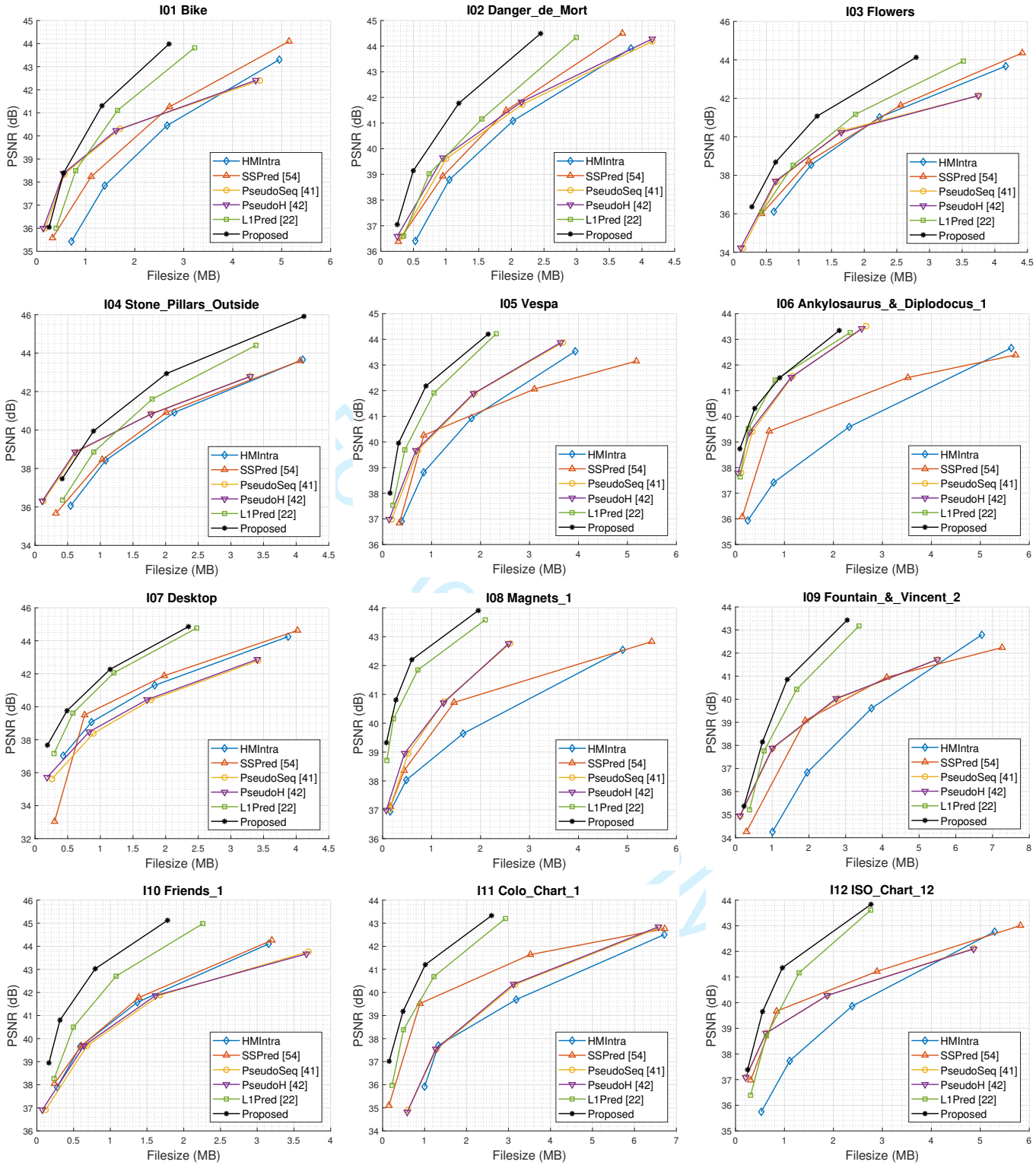

Fig. 8. Rate-distortion curves of the proposed encoding framework and reference algorithms on EPFL dataset of LF images. Images 'House and Lake', 'Palais du Luxembourg', 'Red and White Building', and 'Sophie and Vincent 1' from [53] are used to train the dictionary.

NC1 and NC3 dominate directional intra-prediction, being selected in $60-75\%$ of the cases. One could decide to maintain only NC1 and NC3 to reduce complexity, but that would yield a penalty in terms of coding performance, as in $30-35\%$ of the cases the other directional modes are more efficient in rate-distortion sense.

C. Complexity analysis

The proposed framework was implemented on a machine equipped with an Intel® Xeon® $E5-1650$ v3 3.50GHZ CPU and 64GB of RAM memory, running a 64-bit Windows 8 Operating System. We evaluate the complexity of our proposed

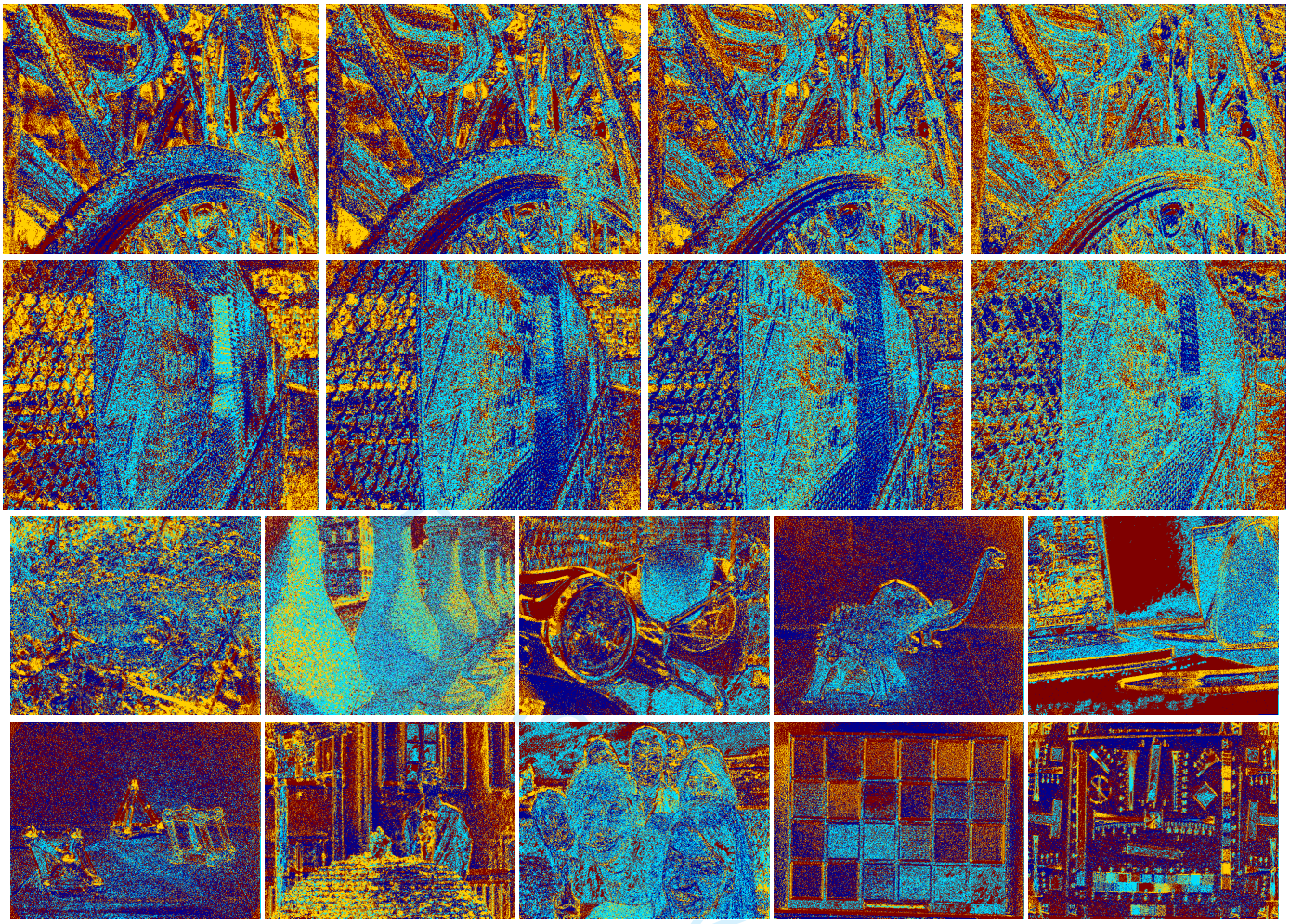

Fig. 9. **(Top)** Mode comparison for $I01$ (1^{st} row) and $I02$ (2^{nd} row), for increasing QP values (from left to right). **(Bottom)** Mode comparison for $I03, I04, I05, I06, I07$ (3^{rd} row) and $I08, I09, I10, I11, I12$ (4^{th} row), for $QP = 22$. The blue colored pixels correspond to macro-pixels encoded with a DL-based intra prediction mode, red to HEVC-based modes, yellow to OP modes, cyan to DP modes (the same color coding as in Fig. 10).

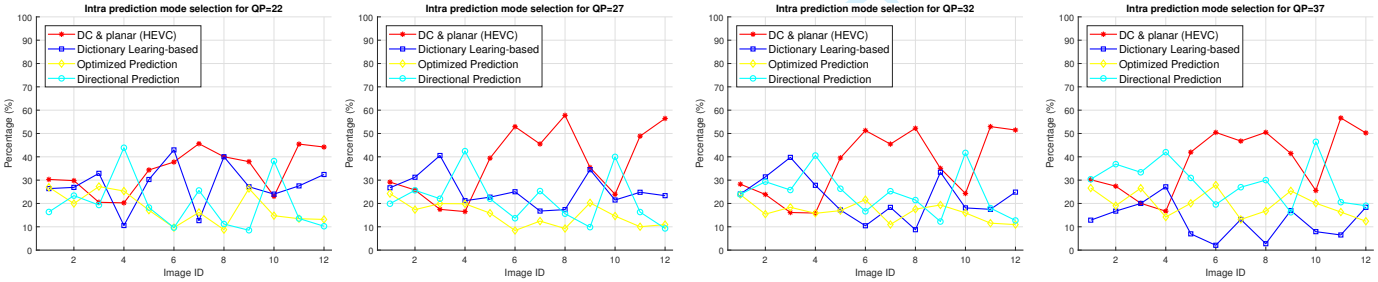

Fig. 10. The selection of the optimal intra prediction mode for the 12 LF images in the EPFL dataset for increasing QP values (from left to right).

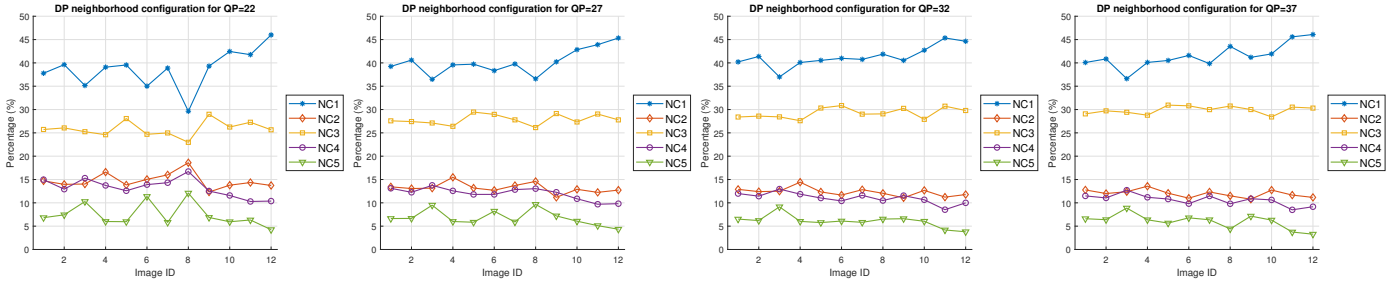

Fig. 11. Intra prediction mode selection between the five NC of the DP, for the EPFL dataset and for increasing QP values (from left to right).

method by the average encoding time for the four tested QPs (22, 27, 32, and 37).

In the HEVC intra prediction mode, the best Coding Unit (CU) is selected from 5 squared macroblock sizes ($64 \times 64, 32 \times 32, 16 \times 16, 8 \times 8, 4 \times 4$) by using RDO. However, the 64×64 macroblock size was not used in our experiments, restricting HEVC to select the best Coding Unit (CU) among $H = 4$ macroblock sizes. For each CU, the runtime spent on RDO, for obtaining the best mode from $k_d = 35$ intra directional predictions is the sum between the time spent on the computation of block-wise distortion and the time spent on bit-cost computation. Hence, the resulting total runtime for HEVC is $k_{HEVC} \approx H \cdot k_d \cdot \gamma_{HEVC} = 140 \cdot \gamma_{HEVC}$, where γ_{HEVC} is a parameter that depends on the method implementation, code optimizations, execution platform, memory allocation of macroblocks of different size, etc.

In the proposed method, the CU is fixed to the size of a macro-pixel. Therefore, macro-pixel prediction is the dominant component in time complexity, which is computed as the sum of complexities of the DL-based prediction, OP prediction, and DP mode prediction. The dictionary is trained off-line, meaning that for the DL-based mode, we only consider the time complexity of macro-pixel prediction and reconstruction, approximately equal to $k_{DL} \cdot \gamma_{DOD}$, where $k_{DL} = 1$ since the method is applied only once, and γ_{DOD} is a parameter which depends on the method implementation, code optimizations, execution platform, memory allocation of macroblocks, macro-pixel size, etc. Due to the distortion optimization among OP modes, matrix multiplication followed by distortion measurement lead to a complexity of approximately $2 \cdot k_{OP} \cdot \gamma_{DOD}$, where $k_{OP} = 32$. The DP mode complexity for one NC is $k_{DP} = 35$, performed for each NC, resulting in a complexity of approximately $5 \cdot k_{DP} \cdot \gamma_{DOD}$. Only two of HEVC's intra-prediction modes (DC and planar) are used in our framework, resulting in a complexity of approximately $k_{DP-HEVC} \cdot \gamma_{DOD}$, where $k_{DP-HEVC} = 2$. Hence, the resulting total runtime of the proposed method is of the order $k_{DOD} \approx (1 + 2 \cdot 32 + 5 \cdot 35 + 2) \cdot \gamma_{DOD} = 242 \cdot \gamma_{DOD}$.

Although rather rudimentary, this complexity analysis indicates that the ratio between the runtime of HEVC and the runtime of the proposed method is approximately $k_{comp} = \frac{k_{HEVC}}{k_{DOD}} = 0.5785 \cdot \frac{\gamma_{HEVC}}{\gamma_{DOD}}$. One can notice that the above time complexity estimation is not taking into account that: **(i)** in HEVC, the runtime of a macroblock may differ for the different sizes of macroblocks; **(ii)** in the proposed method, each type of intra-prediction method can be characterized by a different value of γ_{DOD}.

The experiments have shown that the average runtime on the test set with HEVC is 2876 seconds (around 48 minutes), while the average runtime with the proposed method is 670 seconds (around 11 minutes), indicating that the proposed method is $k_{exp} \approx 4.29$ times faster than HEVC.

V. Conclusions

The paper proposes a novel compression framework for efficient lenslet image coding. Firstly, we introduce a new prediction mode, based on double-sparsity dictionary learning, where each target macro-pixel is represented by a sparse linear combination of dictionary atoms. A novel dictionary generation method accounting for the coding cost of the resulting macro-pixel representation is proposed. A generic dictionary is constructed based on a set of representative LF images which eliminates the need of sending the dictionary to the decoder; only the coefficient values and their locations need to be transmitted, significantly reducing the overall bit cost. Secondly, new optimized linear prediction modes that minimize the residual while accounting for the rate of the resulting macro-pixel representation is proposed. Thirdly, new directional prediction modes for macro-pixels are proposed with the aim of capturing the spatial redundancies by means of directional prediction. Mode selection is controlled by a RDO framework which provides optimal intra coding for each macro-pixel. Experimental results confirm the efficiency of the newly introduced method, as the three proposed intra-coding methods are selected for the large majority of macro-pixels in the lenslet images. The proposed coding system achieves significantly higher PSNR and rate savings compared to reference codecs from the literature, with impressive rate savings going as high as 67.13% and 34.30% against HEVC and the state-of-the-art in lenslet image coding respectively.

Acknowledgment

We would like to thank the authors of [55] for kindly and promptly providing the results of their method.

References

[1] E. Adelson and J. Bergen, "The plenoptic function and the elements of early vision," in *Proc. Comput. Models Vis. Process.* MIT Press, 1991, pp. 3–20.
[2] Y. Taguchi, A. Agrawal, S. Ramalingam, and A. Veeraraghavan, "Axial light field for curved mirrors: Reflect your perspective, widen your view," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, Jun. 2010, pp. 499–506.
[3] B. Wilburnt, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, vol. 2, June 2004, pp. II–294–II–301 Vol.2.
[4] ISO/IEC JTC 1/SC 29/WG 1(ITU-T SG16), Coding of Still Pictures JBIG and JPEG, "JPEG pleno call for proposals on light field coding," https://jpeg.org/downloads/jpegpleno/wg1n73013_pleno_2nd_cfp.pdf.
[5] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007.
[6] R. Ny, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Technical Report*, vol. 2, no. 11, pp. 1–11, 2005.
[7] I. Lytro, "Lytro ILLUM camera," https://www.lytro.com/imaging.
[8] R. GmbH, "3D light field camera solutions," https://www.raytrix.de/produkte.
[9] T. Georgiev, Z. Yu, A. Lumsdaine, and S. Goma, "Lytro camera technology: theory, algorithms, performance analysis," in *Proc. SPIE*, vol. 8667, 2013, p. 86671J.
[10] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Conf. Comput. Graph. Interact. Technol.*, 1996, pp. 31–42.
[11] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar, "Layered 3D: tomographic image synthesis for attenuation-based light field and high dynamic range displays," *ACM Trans. Graph.*, vol. 30, no. 4, p. 95, 2011.
[12] C. Kim, K. Subr, K. Mitchell, A. Sorkine-Hornung, and M. Gross, "Online view sampling for estimating depth from light fields," in *Proc. Int. Conf. Image Process.*, 2015, pp. 1155–1159.

[13] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 569–577, 2003.
[14] F. Zhang, J. Wang, E. Shechtman, Z. Zhou, J. Shi, and S. Hu, "Plenopatch: Patch-based plenoptic image manipulation," *IEEE Trans. Visual. Comput. Graph.*, vol. 23, no. 5, pp. 1561–1573, May 2017.
[15] E. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, 1992.
[16] R. Ng, M. Levoy, M. Bredif, G. Guval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," in *Technical report, Stanford University, Computer Sciences, CSTR*, 2005, pp. 1–11.
[17] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. Int. Conf. on Computational Photography*, April 2009, pp. 1–8.
[18] C. Perwaß and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," in *Proc. SPIE*, vol. 8291, 2012, pp. 8291–8291–15.
[19] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn.*, June 2013, pp. 1027–1034.
[20] DPReview, "LYTRO ILLUM 40 megaray light field camera," https://www.dpreview.com/products/lytro/compacts/lytro_illum/specifications.
[21] R. Zhong, S. Wang, B. Cornelis, Y. Zheng, J. Yuan, and A. Munteanu, "L1-optimized linear prediction for light field image compression," in *Proc. Picture Coding Symp.*, Nuremberg, Germany, 2016, pp. 1–5.
[22] ——, "Efficient directional and l1-optimized intra-prediction for light field images," in *Proc. Int. Conf. Image Process.*, Beijing, China, 2017, pp. 1–5.
[23] G. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
[24] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegands, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.
[25] M. M. A. Endmann and B. Girod, "Progressive compression and rendering of light fields." in *Proc. Int. Symp. Vision Modeli. and Visualiz.*, 2000, pp. 199–204.
[26] X. Dong, D. Qionghan, and Z. Wenli, "Data compression of light field using wavelet packet," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, 2004, pp. 1071–1074.
[27] M. Mazri and A. Aggoun, "Compression of 3D integral images using wavelet decomposition." in *Proc. IEEE Vis. Commun. Image Process.*, 2003, pp. 1181–1192.
[28] A. Aggoun and M. Mazri, "Wavelet-based compression algorithm for still omnidirectional 3d integral images," *Sig. Image and Video Process.*, vol. 2, no. 2, pp. 141–153, 2008.
[29] C. Conti, J. Lino, P. Nunes, L. Soares, and P. Correia, "Improved spatial prediction for 3D holoscopic image and video coding," in *Proc. Eur. Signal Process. Conf.*, 2011, pp. 378–382.
[30] Y. Li, M. Sjostrom, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," in *Proc. IEEE Int. Conf. Acoust Speech Signal Process.*, 2014, pp. 539–543.
[31] L. Lucas, C. Conti, P. Nunes, L. Soares, N. Rodrigues, C. Pagliari, E. da Silva, and S. de Faria, "Locally linear embedding-based prediction for 3d holoscopic image coding using HEVC," in *Proc. Eur. Signal Process. Conf.*, 2014, pp. 11–15.
[32] D. Liu, P. An, R. Ma, C. Yang, and L. Shen, "3D holoscopic image coding scheme using HEVC with gaussian process regression," *Signal Process.: Image Commun.*, vol. 47, pp. 438–451, 2016.
[33] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1308–1319, 2016.
[34] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, 2000.
[35] S. Shi, P. Gioia, and G. Madec, "Efficient compression method for integral images using multi-view video coding," in *Proc. Int. Conf. Image Process.*, 2011, pp. 137–140.
[36] S. Kundu, "Light field compression using homography and 2D warping," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 1349–1352.
[37] S. Adedoyin, W. Fernando, and A. Aggoun, "A joint motion & disparity motion estimation technique for 3D integral video compression using evolutionary strategy," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, 2007.
[38] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, 2006.
[39] M. Rerabek, T. Bruylants, T. Ebrahimi, F. Pereira, and P. Schelkens, "ICME 2016 Grand Challenge: Light-field image compression," *Call for proposals and evaluation procedure*, 2016.
[40] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.
[41] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.
[42] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo sequence based 2-D hierarchical coding structure for light-field image compression," in *Proc. Data Compression Conf.*, 2017, pp. 131–140.
[43] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.
[44] Y. Romano, M. Protter, and M. Elad, "Single image interpolation via adaptive nonlocal sparsity-based modeling," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3085–3098, 2014.
[45] O. Chabiron, F. Malgouyres, J.-Y. Tourneret, and N. Dobigeon, "Toward fast transform learning," *Int. J. Comput. Vision*, vol. 114, no. 2-3, pp. 195–216, 2015.
[46] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, p. 607, 1996.
[47] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Trans. Signal Process.*, vol. 64, no. 12, pp. 3180–3193, 2016.
[48] P. Geladi and B. Kowalski, "Partial least-squares regression: a tutorial," *Analytica chimica acta*, vol. 185, pp. 1–17, 1986.
[49] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, March 1982.
[50] S. Boyd, "L1-norm methods for convex cardinality problems," *Lecture Notes for EE364b, Stanford University. Available at http://www.stanford.edu/class/ee364b*, 2007.
[51] J. Hartigan and M. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
[52] I. Schiopu, *Depth-Map Image Compression Based on Region and Contour Modeling*, ser. Tampere University of Technology. Publication. Tampere University of Technology, 1 2016.
[53] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. Quality of Multimedia Experience*, Lisbon, Portugal, Jun. 2016, pp. 6–8.
[54] 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 *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.
[55] J. Chen, J. Hou, and L. Chau, "Light field compression with disparity-guided sparse coding based on structural key views," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 314–324, Jan. 2018.
[56] M. Rerabek, L. Yuan, L. Authier, and T. Ebrahimi, "ISO/IEC JTC 1/SC 29/WG1 69th meeting EPFL Light-Field Image Dataset," 2015.
[57] J. (JCT-VC), "HEVC reference software, HM version 16.8," https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/.
[58] B. Bross, W.-J. Han, J.-R. Ohm, G. Sullivan, Y.-K. Wang, and T. Wiegand, "High efficiency video coding (HEVC) text specification draft 10," *JCTVC-L1003*, vol. 1, 2013.
[59] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," in *ITU-T Q. 6/SG16 VCEG, 15th Meeting, Austin, Texas, USA*, Apr. 2001.