Steered Mixture-of-Experts for Light Field Video Coding

Vasileios Avramelos[a], Ignace Saenen[a], Ruben Verhack[a,b], Glenn Van Wallendael[a], Peter Lambert[a], and Thomas Sikora[b]

[a]Ghent University - imec - IDLab, Technologiepark-Zwijnaarde 19, 9052 Ghent, Belgium
[b]TU Berlin - Communications Systems Group, Einsteinufer 17, 10587 Berlin, Germany

ABSTRACT

Steered Mixture-of-Experts (SMoE) is a novel framework for representing multidimensional image modalities. In this paper, we propose a coding methodology for SMoE models that is readily extendable to any dimensional SMoE model, thus representing any image modality of any dimension. We evaluate the coding performance of SMoE models of light field video, a 5D image modality, i.e. time, two angular, and two spatial dimensions. The coding consists of the exploiting the redundancy between the parameters of SMoE models, i.e. a set of multivariate Gaussian distributions. We compare the performance of three multi-view HEVC (MV-HEVC) configurations that differ in terms of random access. Each subaperture view from the light field video is interpreted as a single view in MV-HEVC. Experiments validate that excellent coding performance compared to MV-HEVC for low- to midrange bitrates in terms of PSNR and SSIM with bitrate savings up to 75%.

Keywords: Light field video, video coding, multi-view coding, Steered Mixture-of-Experts

1. INTRODUCTION

The consumption of virtual reality (VR) for camera captured content (e.g. 360°video) is lagging behind on the use of VR experiences of computer generated scenes (e.g. in computer games and edutainment software). 360° video allows only rotational head movements around three perpendicular axes for the viewer, but disregards any translational movements in the same 3D coordinate space. To attain the sense of freedom of computer generated VR content, Six Degrees-of-Freedom (6DoF) are required, i.e. three translational movements (walking around and small sideways head movements) combined with three rotational movements (head rotations and tilts). Perceived as a virtual reality by humans when combined, the rendered 2D images are actually processed versions of the higher-dimensional light data that surrounds us. The high-dimensional space is defined by the 5D plenoptic function.[1] However, when there are no occlusions (i.e. "open space" assumption), the 5D space can be reduced to the 4D light field.[2,3] This assumption does not hold for 6DoF in large scenes, however, at the moment this is a widely used simplification.[3]

One promising novel methodology that aims for the representation of multidimensional image modalities has been introduced, namely *Steered Mixture-of-Experts* (SMoE). It directly models the underlying plenoptic function in a continuous, analytical form, or a lower-dimensional projection of this function.[1] It does so by identifying coherent regions in the coordinate space of the samples and optimizes local linear regressors for that segment in the coordinate space. The total regression corresponds to a smoothed piecewise linear approximation of the plenoptic function (or of a lower-dimensional projection). Currently, SMoE has been successfully applied for images, video, and static 4D light fields for coding, with competitive rate-distortion (RD) results for low- to mid-level bitrates.[4–6] The local regressors currently reported are only linear and thus modeling very high spatial frequencies is challenging, however, the theory does not limit the nature of the local regressors and further developments are an active area of research. Recently, work has been published on the modeling and rendering of SMoE models for light field video, however, without coding.[7,8] The goal of this paper is to revise and generalize previous SMoE coding methods and compare with the state-of-the-art in terms of RD-performance.

Evidently, SMoE has proven its potential by presenting beneficial properties for the distribution of 6DoF visual content compared to traditional image coding methods.[7–11] For instance, in the case of rendering it has

Further author information:
E-mail: vasileios.avramelos@ugent.be

three important properties. Firstly, view-rendering is very lightweight and pixels are coded independent from one another. Secondly, SMoE is a space-continuous representation, thus rendering at arbitrary resolution consists of merely sampling this function. Finally, all local light information in a certain point in the physical space is also localized in the SMoE model.

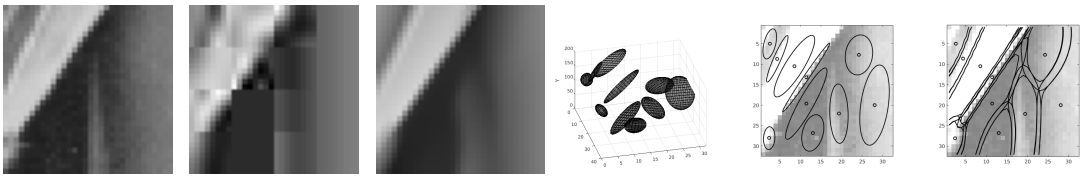
Digital image and video compression techniques have been an important field of research since the 1950s. Standardized image and video coders typically rely on a transform step (e.g. wavelet or DCT) and *Differential Pulse-Code Modulation* (DPCM) (e.g. intra-prediction, and motion compensation). As a result, the current state-of-the-art coders like *High Efficiency Video Coding (HEVC)* are based on hybrid transform/DPCM coding schemes which consist mainly of the above mentioned techniques.[12] The serial nature of these old paradigms (e.g., intra-prediction) makes it impossible to really achieve pixel-level parallelism which is more and more desirable for modern hardware architectures. Furthermore, traditional coding schemes based on dense sample/coefficient grids do not scale easily towards higher dimensional image modalities. Each dimension that is added (e.g. time dimension in video, or two angular dimensions in 4D light fields) lets the amount of samples to be stored grow exponentially with the dimensionality of the image modality. Multi-view HEVC (MV-HEVC) has been introduced as an HEVC extension for coding 3D and multi-view video.[13] There have been many multi-view video coding research efforts for trying to efficiently code light field video. Conventionally, multi-view coding (MVC) is using the MV-HEVC extension for coding light field video by treating each subaperture view as a different video sequence. On the other hand, multi-view coding plus depth (MVD) is using the 3D-HEVC extension for coding light field video by treating a central view as a video sequence and the rest subaperture views as disparity maps (depth maps).[14]

The *Moving Picture Experts Group* (MPEG) has started efforts to standardize a 6DoF video format by the year 2021.[15] They aim at a process with two phases: (1) identifying the most important 2D views, and (2) rely on view synthesis methods to render other 2D views at decoder side. The identified views are expected to be coded using the same hybrid DPCM/transform coding approaches.[16] We claim that there are two main concerns. First, the view synthesis may require considerable computational complexity at the decoder side. Secondly, the serial nature of the paradigms is far from optimal as the prediction order is much less evident. In video coding, frames can be buffered if a logical order exists between them, e.g. when the frames are time-consecutive. However, considering the freedom to select a particular point of view in a 6DoF VR experience, no such logical order exists. As such, buffering and differential coding become more challenging.

2. STEERED MIXTURE-OF-EXPERTS FOR LIGHT FIELD VIDEO

Steered Mixture-of-Experts (SMoE) is a novel framework for approximation of image modalities with many applications, such as image modality coding, scale conversion (e.g. frame interpolation), and image description (e.g. depth estimation). An in-depth overview about SMoE for images and static 4D light fields is presented in Verhack et al.[11] Due to the sparse structure in SMoE, it is readily extensible towards higher dimensional image modalities, such as 6DoF content. This is in stark contrast to traditional image coding schemes which rely on dense sample-grid structures. Moreover, SMoE departs significantly from the conventional coding methods by operating in the spatial domain and thus not using any kind of transform coding. Instead of storing exactly the samples or the transform coefficients that define the image, this method relies on modeling the underlying generative function that could have given rise to the samples. Generally, this underlying function corresponds to lower-dimensional projections of the plenoptic function.[1]

The function approximation of the underlying generative function is done by identifying coherent, stationary regions in the image modality. Each segment is modeled using a single N-dimensional entity, which we call a *kernel* or *component*. SMoE is based on the divide-and-conquer principle that is present in all *Mixture-of-Experts* (MoE) approaches. These methods are well-known in machine learning.[17] The input space (in SMoE this is the *coordinate space*) is divided in multidimensional soft-segments/hyper-volumes using a gating function. For images, such a segment is a (irregular) patch of pixels, for video this is a patch of pixels along frames. Analoguously, for light fields, these are multidimensional patches of pixels along angular dimensions. Local regressors (or *experts*) are sought that locally approximate the function optimally. The gating function then lets experts collaborate in segments where they are trustworthy, i.e. these segments can overlap in areas.



(a) Original (b) JPEG (c) SMoE (d) 3D GMM (e) Topview (f) Softmax

Figure 1. An example of the modeling with 10 components and reconstruction of a 32x32 pixel crop from *Lena* (1(a)). For a grayscale image, the coordinate space X is 2D and the colorspace Y is 1D. Modeling the joint probability function of both X and Y using a Gaussian Mixture Model results in 3D Gaussian kernels (1(d)). Each kernel thus defines a 2D gradient as the *expert* function ($X \mapsto Y$). The gating function is defined by the soft-segmentation (1(f)). Both JPEG (1(b)) and SMoE (1(c)) are coded at 0.35 bpp.[4]

SMoE is based on the Bayesian, or "alternative" definition of the Mixture-of-Experts model.[17] The Bayesian Mixture-of-Experts approach jointly models the joint probability of the input space X and the output space Y using a *Gaussian Mixture Model* (GMM). Each Gaussian kernel then simultaneously defines the gating function (soft-segmentation of X) and the local regressors (through the conditional probability function $Y|X$). Although recently, approaches have been proposed to separate the optimization of the gating and expert functions. In such cases, the parametrization is not limited to GMMs.[18] However, in this work, we assume the model takes on the form of a GMM as it has been in all other cases.

In SMoE, where the input space is the *coordinate space* X (i.e. sample locations) and the output space is the *color space* Y (i.e. sample amplitudes), one such Gaussian then corresponds to one kernel as mentioned above. The gating function is thus defined by the probability that a coordinate belongs to a Gaussian, and each Gaussian simultaneously defines an expert function, namely the conditional color amplitudes, given a coordinate. In general, the SMoE allows to query the model at any sub-pixel coordinate to yield the most optimal amplitude in a Bayesian interpretation.

SMoE thus arrives at a sparse representation. The whole image modality is represented as a set of Gaussian kernels. These kernels are defined by their centers and their steering parameters. The coordinate space is 2D, 3D, 4D, and 5D in the case of respectively images, video, static and dynamic light fields.[4–6] The color space for color images is conventionally represented as a 3D space, e.g. RGB or YCbCr. As the Gaussians model the joint probability of the coordinate and color space, we thus arrive at respectively 5D, 6D, 7D, and 8D Gaussian kernels. The parameters of these kernels are typically estimated using computational efficient variations of the *Expectation-Maximization* (EM) algorithm.[19] Due to this likelihood optimization, kernels will steer along the dimensions of the highest correlation, e.g., along spatial or temporal consistencies. Very promising MSE-based modeling approaches to find the kernel parameters have been introduced recently.[18,20] In this paper, however, we work on likelihood optimized SMoE models without loss of generality.

Fig. 1 shows an example of the compression capability of the SMoE approach for coding a 32x32 pixel crop of Lena at 0.35 bits/sample in comparison to JPEG at same rate. Clearly, the edges are reconstructed with convincing quality and sharpness, using merely 10 components.[4] In general, the framework achieves good performance for low-to-mid bitrates compared to the state-of-the-art, which is considerable taking into account the high difference in maturity. Fig. 2 illustrates a SMoE light field reconstruction using only 8960 kernels.[6]

Fig. 3 illustrates the high-level coding process. The encoding step thus relies on an iterative optimization process similar to other machine learning approaches. Due to the specific structure of the data in image modalities, many heuristics can be used to arrive at an efficient modeling scheme. The focus of this paper is the coding of the model parameters. The parameters are decorrelated, quantized and further binarized by using an arithmetic coder.[6,11] For the remainder of this paper, we will provide the revised coding methodology applied on SMoE models of light field video, without loss of generality. Finally, we will compare our proposed scheme with MV-HEVC. For a more detailed elaboration on the mathematics and theory present in SMoE which are needed for the modeling and reconstruction of the views from SMoE models, we refer the reader to the SMoE work on 4D light fields.[11]

(a) Original (b) SMoE

Figure 2. *Bikes*[21,22] light field example (K=8960), showing a central view with $(a_1, a_2) = (7, 7)$. The original light field has $15x15x626x434$ samples. Consequently, each Gaussian kernel "covers" 6822 samples on average. (Mean $\text{PSNR}_{\text{YCbCr}}$: 30.71 dB, mean SSIM_{Y}: 0.86, evaluation as in[21]).

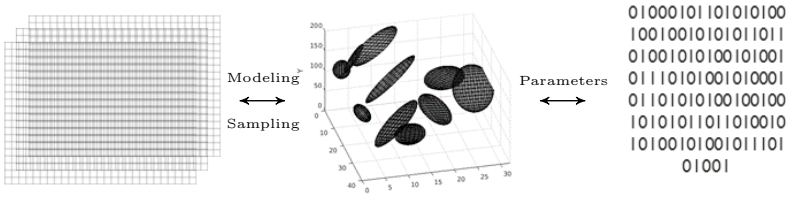


Figure 3. A high-level view of the encoding scheme based on SMoE. Acquired sample grids are being modeled into a set of Gaussian kernels. In order to store this model only the parameters of these Gaussians need to be binarized. Decoding then consists of unpacking the Gaussian parameters and rendering the desired view.

3. CODING OF SMOE MODELS

In this section, we will limit ourselves to the specific case of light field videos without loss of generality. Our coordinate space X is thus 5D (time, two angular, and two spatial dimensions) and the color space is 3D (YCbCr). Assume training data $D = \{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=1}^N$, with $x \in X$ and $y \in Y$. In order to avoid confusion , we will use the notation $\mathcal{Y}$ for the luma channel and Y for the output space. The joint probability density of coordinates and amplitudes is modeled using a GMM with K kernels as follows:

$$p_{XY}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\mu}_j, R_j) \tag{1}$$

with $\sum_{j=1}^{K} \pi_j = 1, \boldsymbol{\mu}_j = \begin{bmatrix} \boldsymbol{\mu}_{X_j} \\ \boldsymbol{\mu}_{Y_j} \end{bmatrix}, R_j = \begin{bmatrix} R_{X_jX_j} & R_{X_jY_j} \\ R_{Y_jX_j} & R_{Y_jY_j} \end{bmatrix}$, and $\mathcal{N}(\cdot)$ being the multivariate normal distribution.

The parameters of this model are thus $\Theta = [\theta_1, \cdots, \theta_K]$, with $\theta_j = (\pi_j, \boldsymbol{\mu}_j, R_j)$, respectively being the priors, centers, and covariances for each multivariate Gaussian kernel. However, not all of these parameters are being coded as some are of no importance or off less importance. More specifically, we do not encode the 3×3 color covariance matrix $R_{Y_jY_j}$ as it is not necessary for the reconstruction, and we also exclude $R_{X_jY_j^{\text{Cb}}}$ and $R_{X_jY_j^{\text{Cr}}}$, i.e. the covariance of coordinates and chroma amplitudes Cb and Cr. As such, the color gradients are assumed to be constant. We leave this out as the human visual system is less susceptible to changes in chroma values. Each kernel at decoding side thus only has mean color value, but no color gradient. In previous works, the priors π, had always been neglected and we assumed uniform priors at decoding side, i.e. $\pi_j = 1/K$. However, in this work, we have included the priors into the bitstream as they do have significant impact on the reconstruction quality.

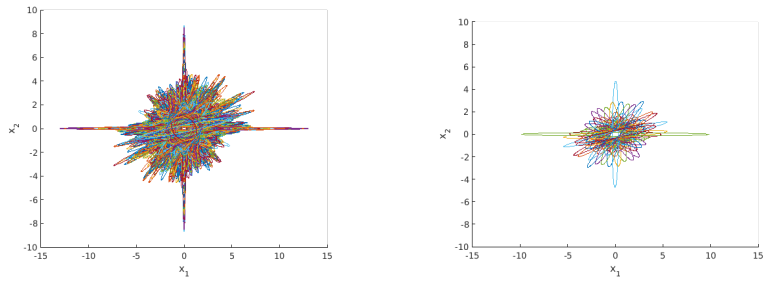


Figure 4. Codebook of size 64 (right) made from a set of 8960 2-D normalized covariances (left)[11]

The remaining parameters to be coded into the bitstream thus consist of: $(\pi_j, \boldsymbol{\mu}_j, R_{X_jX_j}, R_{X_jY_j^{\mathcal{Y}}})$, i.e. respectively the priors (scalar), the centers (5+3=8D), the covariance matrix in the coordinate space (5×5) and the covariance values of the coordinate space and the luma ($\mathcal{Y}$) amplitudes (5D).

The main encoding strategy goes as follows:

1. For $\{R_{X_jX_j}\}_{j=1}^{K}$:
   (a) Normalize all $\{R_{X_jX_j}\}_{j=1}^{K}$ to $R_{X_jX_j} = s_j \tilde{R}_{X_jX_j}$ with $\det(\tilde{R}_{X_jX_j}) = 1$ and encode s_j (determinant of the original covariance) per kernel in the next step
   (b) Build dictionary of $L(< K)$ normalized covariance matrices $\{C_{X_jX_j}\}_{i=1}^{L}$
   (c) Encode dictionary separately analogously to (2.c)-(2.e)
2. For $\{(\pi_j, \boldsymbol{\mu}_j, R_{X_jY_j^{\mathcal{Y}}}, s_j)\}_{j=1}^{K}$:
   (a) Sort kernels based on a greedy fashion along centers μ_j in order to minimize the distance between two consecutive kernel centers
   (b) Transform parameters to be laplacian distributed per parameter
   (c) Merge all distributions on the same laplacian distribution with variances according to the importance of the parameter
   (d) Quantize the values uniformly
   (e) Encode using a laplacian adaptive arithmetic coder

3.1 Coordinate covariance $R_{X_jX_j}$ quantization

As in related work,[6,11] we employ a vector quantization-like method for coding the window covariance $R_{X_jX_j}$. We propose a minibatch EM-like algorithm based on the *Kullback-Leibler* (KL) divergence.[11] As such, the probability densities are compared, which are more informative than the covariance parameters. We normalize all $R_{X_jX_j}$ by $|R_{X_jX_j}|^{(1/d)}$. In the case of $R_{X_jX_j}$ for light field video, d equals 5. As such, the constructed codebook contains normalized shapes with a determinant of one. The coding of the magnitude of the shape, i.e. $s_j = |R_{X_jX_j}|^{(1/d)}$ is discussed in the next subsection.

The KL-divergence of two multivariate Gaussian distributions $P \sim \mathcal{N}(\boldsymbol{\mu}_P, R_P)$ and $Q \sim \mathcal{N}(\boldsymbol{\mu}_Q, R_Q)$ is given by

$$\mathrm{D}_{\mathrm{KL}}(P \parallel Q) = \frac{1}{2}\left[\log\left(\frac{|R_P|}{|R_Q|}\right) - d + \mathrm{trace}(R_Q^{-1}R_P)\right] + \frac{1}{2}\left[(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^T R_Q^{-1}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)\right] \tag{2}$$

As our data is normalized, $|R_P|$ and $|R_Q|$ equal one. Furthermore, the windows are assumed to be centered on the origin, i.e. $\boldsymbol{\mu}_P$ and $\boldsymbol{\mu}_Q$ are zero. In order to obtain a symmetric similarity measure, we define our distance as

$$\begin{aligned}\mathrm{dist}(P,Q) &= \frac{\mathrm{D}_{\mathrm{KL}}(P \parallel Q) + \mathrm{D}_{\mathrm{KL}}(Q \parallel P)}{2} \\ &= \frac{1}{4}\left(-2d + \mathrm{trace}(R_Q^{-1}R_P) + \mathrm{trace}(R_P^{-1}R_Q)\right)\end{aligned}$$

Covariances are clustered around a centroid using $\text{dist}(P, Q)$ and at each iteration the new centroid covariance C_l is calculated as the mean covariance of the members of the cluster l and renormalized. Fig. 4 illustrates the algorithm on a 2-D dataset.

This codebook is trained at encoder side, and transformed to ensure robustness. As each C_l is semi-positive definite, C_l can be decomposed using Cholesky: $C_l = A^T A$. A is vectorized into $\boldsymbol{a}$ of length 15 in the case of light field video. We thus arrive at a matrix of size $L \times 15$. Next, we normalize all columns to have mean zero and variance one. The real variances and means are sent as header information. Each column now is assumed to follow the same distribution. We further quantize the values uniformly into b_{book} bits according to the limits of each column. The limits and b_{book} are transmitted as header info. We assume the distribution to behave laplacian. Using a laplacian adaptive arithmetic coder (as detailed in the next section), we entropy encode the dictionary. Even though the laplacian assumption is not strictly true (some columns are only half-laplacian), we found that it works well in practice. The resulting bits-per-symbol converges close to the entropy. Finally, at decoder side, the multiplication $A^T A$ ensures the reconstructed covariance to be semi-positive definite again.

3.2 Kernel parameters $(\pi_j, \boldsymbol{\mu}_j, R_{X_j Y_j^{\mathcal{Y}}}, s_j)$ quantization and arithmetic coding

3.2.1 Sorting and transformation to Laplace distributions

The centers $\boldsymbol{\mu} = [\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y]$ are difference coded by defining a path that comprises every component exactly once in a greedy fashion. Start with the component j closest to $(0, 0)$. Find component $k, (k \neq j)$, so that $|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k|$ is minimal. Because of the prediction from the previous kernel center, we obtain that the new transformed $\tilde{\boldsymbol{\mu}}_j^k$ are Laplacian-distributed centered around zero. Next, sort all other kernel parameters $(\pi_j, R_{X_j Y_j^{\mathcal{Y}}}, s_j)$ using the same permutation.

The priors π_j are not Laplacian-distributed, however, they can be easily transformed to have a Laplace distribution. First, shift the mode onto zero. Secondly, negate every second value in order to have symmetry around zero. Consequently, we arrive at a transformed $\tilde{\pi}_j$ which has a Laplacian distribution. The values of $R_{X_j Y_j^{\mathcal{Y}}}$ and s_j are relatively close to being Laplacian-distributed from the start and are thus not further processed. Each kernel's $R_{X_j X_j}$ is matched to the covariance dictionary and it's index is saved. These indexes are uniformly distributed as the dictionary method is a form of non-linear quantization, and are thus coded without arithmetic coding.

Finally, we can conclude that the resulting coefficients $(\tilde{\pi}_j, \tilde{\boldsymbol{\mu}}_j, R_{X_j Y_j^{\mathcal{Y}}}, s_j)$ are Laplacian-distributed per parameter and are then further processed for arithmetic coding according to the desired precision for each type of coefficient as follows.

3.2.2 Distribution merging, quantization and arithmetic coding

The transformed priors $\tilde{\pi}_j$, the 8D differenced centers $\tilde{\boldsymbol{\mu}}_j$, the 5 dimensions in $R_{X_j Y_j^{\mathcal{Y}}}$, and the shape magnitude s_j are concatenated in one $1 + 8 + 5 + 1 = 15$D vector s per kernel. We arrive at a $K \times 15$ matrix of kernel parameters. There values are further normalized per column as follows

$$\tilde{s}_j^i = \frac{s_j^i - E[s^i]}{c^i \sigma_{s^i}} \tag{3}$$

with $c^i \geq 1$ being the ratio determining how much more subsampled the coefficient i needs to be compared to the spatial coordinates of the kernel. We thus set $c^4 = c^5 = 1$ as the baseline, i.e. c^i with $i \notin \{4, 5\}$ determines how much less important coefficient i is compared to the spatial location center (μ_j^4, μ_j^5). Hereby we assume that the precision of the location always will be the highest compared to the other coefficients. Consequently, the distribution of the coefficient i with $c^i > 1$ is squeezed together, resulting in less bits being spent on these parameters. Finally, we vectorize the matrix column-wise into one stream of $15 \times K$ symbols.

Next, quantization is performed uniformly based on the limits of s. As such, we are able to combine different quantization steps for each coefficient, while still using a single arithmetic coder. The same laplacian adaptive arithmetic coder is employed as in.[23] Hereby, only the mean and variance of the source needs to be transmitted, as such, we do not need to transmit the full probability per symbol. Finally, after each symbol that has been processed, the distribution is adapted both at encoder and decoder side.

4. EXPERIMENTS

4.1 Setup

For the three light field sequences *cats*, *train1* and *train2*,[24] we evaluated the coding performance of three different MV-HEVC scenarios and compared them to the SMoE light field video coding approach described in Sec. 3. The dataset consists of two light field video sequences of resolution 512×352 and one at 544×320 at 30 fps for approximately 100 frames and 8×8 views. Note that the sequences originate from interpolating light fields originating from a plenoptic camera and include some artifacts from this process.[24] In Sec. 4.2, a detailed description of the MV-HEVC configuration used in this work is given.

Figure 5. The three different light field video sequences used in this work. From left to right: *cats* - 512×352 109 frames, *train1* - 512×352 84 frames, and *train2* - 544×320 97 frames.

In order to assess the objective quality of the whole light field video, all views at every frame should be assessed. However, in order to speed up the measurements, we choose one view per frame and iterate over five possible views. The five randomly chosen views are *view(2,2)*, *view(3,6)*, *view(4,4)*, *view(6,7)* and *view(7,2)*, where for instance *view(2,2)* corresponds to view with $[X, Y]$ coordinates equal to $[2, 2]$.

4.2 MV-HEVC

For comparison we used the MV-HEVC reference software HM-16.5,[13,25] for encoding and decoding the three different light field video sequences in three different scenarios. In the first scenario, we investigated the case of independently encoding every frame and every view (MVC-allIntra) for enabling random access in time and space. In the second scenario, we enable inter-frame prediction but not inter-view prediction (MVC-interFrame) i.e, for every different light field video view, subsequent frames can be predicted from previous or following frames but they cannot use frames from other views as a reference. In the last scenario, we enable both inter-view and inter-frame prediction (MVC-full) to present a more efficient fully-referenced light field coding scheme but without random access capabilities.

For enabling dependencies between views, we allowed every view to be predicted by two other already coded views. First of all, we indexed the views in the 2D space by using a rather modified spiral scanning technique as shown in Fig. 6 (left). Starting from the corners of the grid, we then scan the 2D space by revolving in a clockwise way until we reach the center viewpoints. Obviously, the number of possible matrix scanning topologies is unlimited, and other techniques such as raster scan or Zig-zag scan can also be used. However, as seen in Wang et al, by increasing the number of bi-directional dependencies in the encoding structure, the bitrate can be reduced and on top of that, vertical correlations should also be considered for a more optimal prediction structure.[26] Therefore, in this work, bi-directional referenced frames in time, and views in space, have been maximized for comparing the SMoE coder to a more efficient MV-HEVC light field coding scheme. There is a vast amount of possibilities to choose which view can be referenced from which two other views. In this work, we choose to start the coding process by first encoding the corner views of the light field view grid, and as we follow the modified reverse spiral structure, we choose to always reference the closest horizontal, vertical or diagonal views by slightly preferring the horizontal domain since the human visual system is more biased to horizontal correlations.[27] An example of the inter-view prediction structure for a downscaled 4×4 views case is shown in Fig. 6 (right).

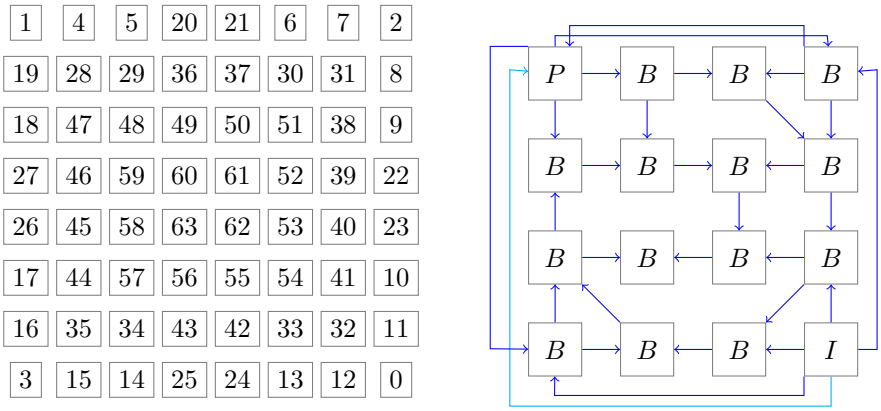


Figure 6. Grid view of the 8×8 light field view ordering structure (left) and inter-view prediction structure for a 4×4 views case (right). Similar referencing structure has been derived for the 8×8 case.

4.3 Results

In this section we present the comparison results for light field video coding between the MV-HEVC configured as in Sec. 4.2, and the presented light field video coding framework as described in Sec. 2. Bitrates were calculated as the total file size divided by the duration of the sequence. We compared bitrate versus video quality and therefore we present RD-curves by using Peak Signal to Noise Ratio (PSNR) and the Structural Similarity (SSIM) index as the video quality assessment metrics for this work. However, we believe that a subjective quality assessment would be useful in this work as well, since it is not yet clear which objective video quality metrics are best fitted for evaluating light field video. Therefore, in Fig. 8 we present an example of a subjective comparison between SMoE and the best inter-predicted MV-HEVC configuration used in this work. As it can be observed, the spatial artifacts introduced by SMoE, such as blurriness (spatial and temporal), are more visible, while for MV-HEVC blocking artifacts are more present than blurriness. Note here, that temporal blocking artifacts can also be clearly observed during video playback for the MV-HEVC case, whereas SMoE has strong view consistency over angular and temporal dimensions. Superior view consistency compared to MV-HEVC along angular dimensions for static light fields has been shown in previous work.[11]

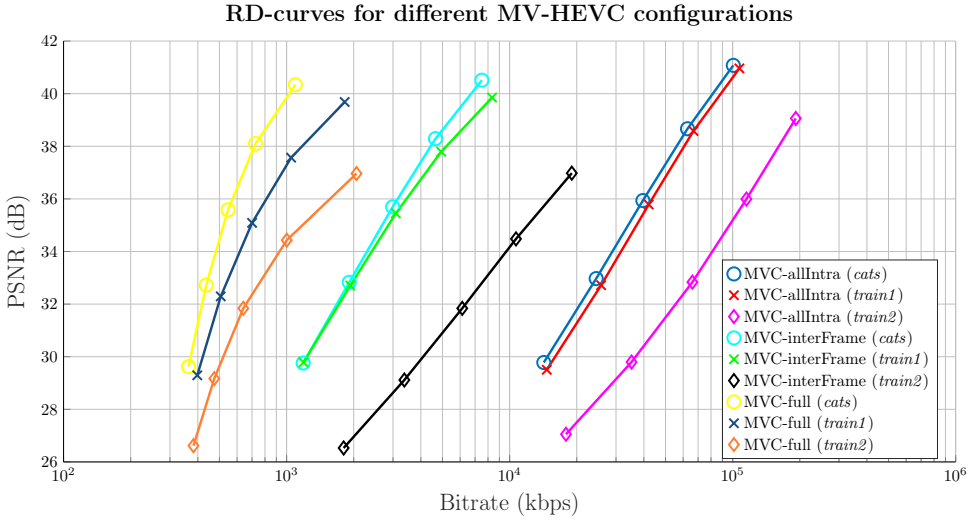
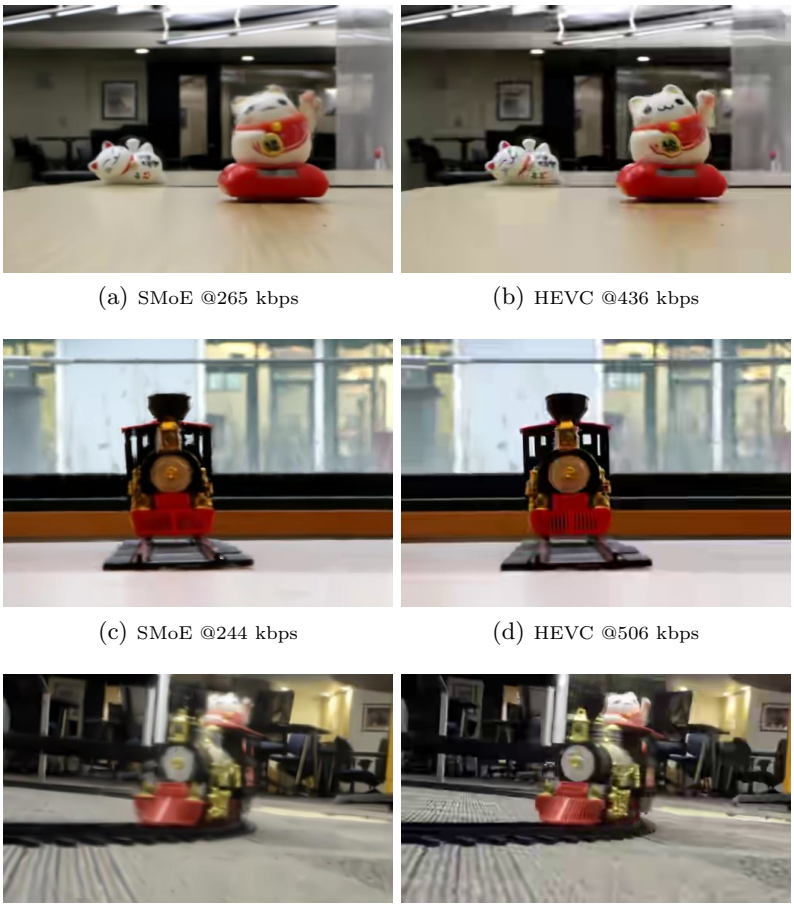


Figure 7. RD-curves for the three datasets *cats*, *train1* and *train2* when coded with different MV-HEVC configurations. For better visualization, logarithmic scale has been used. It can be observed, that by adding one level of random accessibility to our configuration, the bitrate increases by a factor of ×10.



(a) SMoE @265 kbps (b) HEVC @436 kbps

(c) SMoE @244 kbps (d) HEVC @506 kbps

(e) SMoE @227 kbps (f) HEVC @474 kbps

Figure 8. Subjective comparison of SMoE (left) vs. MV-HEVC (right) used in this work for datasets *cats* (top), *train1* (middle) and *train2* (bottom) both at 33dB, 32dB and 30dB PSNR respectively. However, the bitrate savings for this example are in the range of 50% for SMoE.

In Fig. 7, it is shown the RD-performance of the MVC-allIntra and MVC-interFrame scenarios (see Sec. 4.2) for three different light field video sequences. In these different random access scenarios, MV-HEVC performs well as expected, however it ranges in higher bitrates when compared to MVC-full and SMoE. Therefore, we only present a comparison of MVC-full and SMoE in Fig. 9, where for every test sequence the RD-curves are presented. As mentioned in Sec. 1, SMoE is competitive at low- to mid- bitrates and that can be seen in our end results (Fig. 9). However, while for MV-HEVC the bitrate and video quality increases linearly, for SMoE there is an early cut-off point in terms of quality at 0.955, 0.95 and 0.85 for SSIM, and 35.5dB, 34dB, 32dB PSNR for *cats*, *train1* and *train2* respectively.

5. CONCLUSIONS

In this paper we have proposed a coding scheme for SMoE models by removing redundancy between the Gaussian kernel parameters and by using arithmetic coding on the transformed parameters. Furthermore, we have experimentally evaluated the efficiency of the proposed scheme on SMoE models that represent light field videos, i.e. having a 5D coordinate space and a 3D YCbCr color space. We used MV-HEVC for light field video coding as the state-of-the-art reference. Results showed that for low- to mid-range bitrates, SMoE can outperform MV-HEVC with only a single I-frame (i.e. with low random access capabilities) in terms of PSNR and SSIM with bitrate savings of up to 75% on a dataset of three short-baselined light field videos.

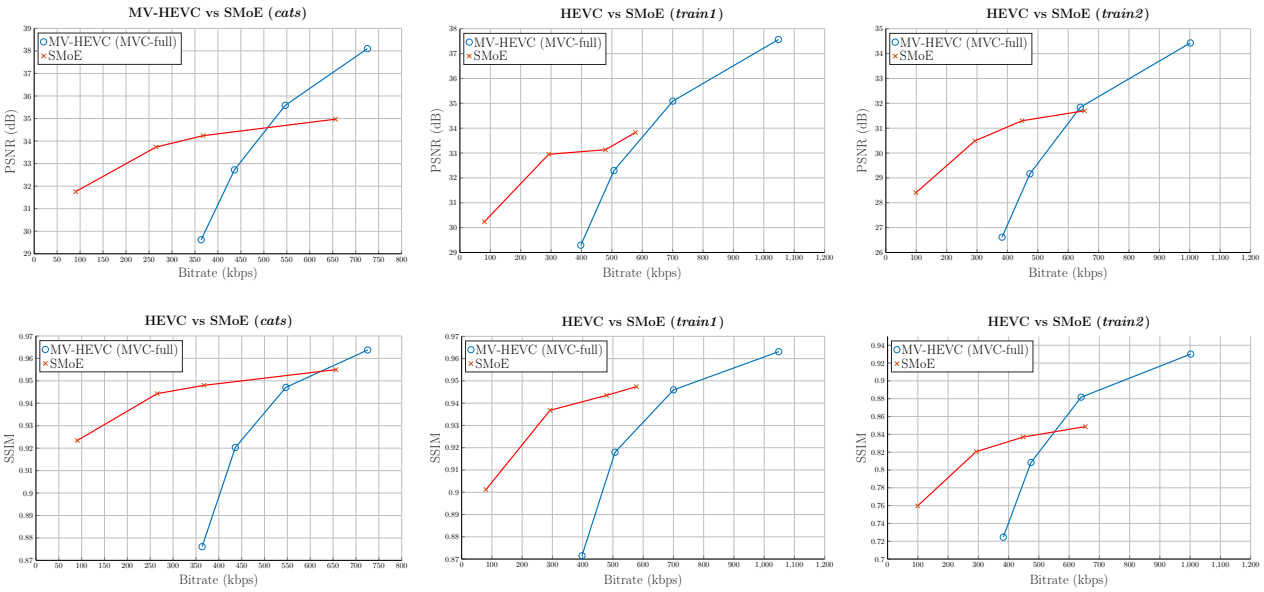


Figure 9. Rate-distortion performance of MV-HEVC versus SMoE for three different light field video sequences. For MV-HEVC the MVC-full configuration has been used, which exploits both inter-frame and inter-view dependencies. For low- to mid-bitrates SMoE can already reach acceptable objective quality in terms of PSNR and SSIM.

Apart from coding gain, SMoE models offer a number of additional features such as pixel-level parallel decoding, and granular random access (after decoding the kernel parameters), however, this has been less the focus of this work and remain to be further investigated and compared. Further optimization of the SMoE modeling approach (such as MSE-optimization) and coding techniques is ongoing research. Additionally, further experimental evaluation on longer sequences or other types of light fields (e.g. more wide-baselined) remains to be explored. Finally, light field video compression and especially the evaluation of the performance of such algorithms in terms of quality and view consistency is not well understood at this point and remains an active research track. However, it is clear that the SMoE framework and associated coding methods are very promising representations for higher dimensional image modalities.

Acknowledgments

The research activities described in this paper were funded by IDLab (Ghent University - imec), Communication Systems Lab (Technische Universität Berlin), Flanders Innovation & Entrepreneurship (VLAIO), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union. The computational resources (STEVIN Supercomputer Infrastructure) and services used in this work were kindly provided by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation, and the Flemish Government department EW.

REFERENCES

[1] Adelson, E. and Bergen, J., "The plenoptic function and the elements of early vision," *Computational Models of Visual Processing (MIT Press)* , 3–20 (1991).

[2] Levoy, M. and Hanrahan, P., "Light field rendering," in [*Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '96*], 31–42, ACM Press, New York, New York, USA (1996).

[3] Wu, G., Masia, B., Jarabo, A., Zhang, Y., Wang, L., Dai, Q., Chai, T., and Liu, Y., "Light Field Image Processing: An Overview," *IEEE Journal of Selected Topics in Signal Processing* , 1–1 (2017).

[4] Verhack, R., Sikora, T., Lange, L., Van Wallendael, G., and Lambert, P., "A universal image coding approach using sparse mixture-of-experts regression," in [*Proceedings of the IEEE International Conference on Image Processing (ICIP)*], 2142–2146 (2016).

[5] Lange, L., Verhack, R., and Sikora, T., "Video representation and coding using a sparse steered mixture-of-experts network," in [*Picture Coding Symposium (PCS)*], 1–5 (2016).

[6] Verhack, R., Sikora, T., Lange, L., Jongebloed, R., Van Wallendael, G., and Lambert, P., "Steered mixture-of-experts for light field coding, depth estimation, and processing," in [*Proceedings of the IEEE Conference on Multimedia and Expo (ICME)*], 1183–1188 (2017).

[7] Verhack, R., Van Wallendael, G., Courteaux, M., Lambert, P., and Sikora, T., "Progressive Modeling of Steered Mixture-of-Experts for Light Field Video Approximation," in [*Picture Coding Symposium '18*], (2018).

[8] Saenen, I., Verhack, R., Avramelos, V., Van Wallendael, G., and Lambert, P., "Hard Real-Time, Pixel-Parallel Rendering of Light Field Videos using Steered Mixture-of-Experts," in [*Picture Coding Symposium '18*], (2018).

[9] Verhack, R., Madhu, N., Van Wallendael, G., Lambert, P., and Sikora, T., "Steered Mixture-of-Experts Approximation of Spherical Image Data," in [*Accepted for publication at EUSIPCO '18*], (2018).

[10] Avramelos, V., Verhack, R., Saenen, I., Van Wallendael, G., Goossens, B., and Lambert, P., "Highly Parallel Steered Mixture-of-Experts Rendering at Pixel-Level for Image and Light Field Data," *Submitted to Journal on Real-Time Image Processing* (2018).

[11] Verhack, R., Sikora, T., Lange, L., Van Wallendael, G., and Lambert, P., "Steered Mixture-of-Experts for 4-D Light Field Approximation, Coding, and Description," *Submitted to IEEE Transactions on Multimedia* (2018).

[12] Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T., "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology* **22**, 1649–1668 (dec 2012).

[13] Hannuksela, M. M., Yan, Y., Huang, X., and Li, H., "Overview of the multiview high efficiency video coding (mv-hevc) standard," *2015 IEEE International Conference on Image Processing (ICIP)* , 2154–2158 (2015).

[14] Conti, C., Kovács, P. T., Balogh, T., Nunes, P., and Soares, L. D., "Light-field video coding using geometry-based disparity compensation," *2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)* , 1–4 (2014).

[15] Domanski, M., Stankiewicz, O., Wegner, K., and Grajek, T., "Immersive visual media MPEG-I: 360 video, virtual navigation and beyond," in [*2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*], 1–9, IEEE (may 2017).

[16] Hinds, A. T., Doyen, D., and Carballeira, P., "Toward the realization of six degrees-of-freedom with compressed light fields," in [*2017 IEEE International Conference on Multimedia and Expo (ICME)*], 1171–1176, IEEE (jul 2017).

[17] Yuksel, S. E., Wilson, J. N., and Gader, P. D., "Twenty Years of Mixture of Experts," *IEEE Transactions on Neural Networks and Learning Systems* **23**, 1177–1193 (aug 2012).

[18] Tok, M., Jongebloed, R., Lange, L., Bochinski, E., and Sikora, T., "An MSE approach for training and coding steered Mixtures of Experts," in [*Picture Coding Symposium (PCS '18)*], (2018).

[19] Moon, T., "The Expectation-Maximization Algorithm," *IEEE Signal Processing Magazine* **13**(6), 47–60 (1996).

[20] Bochinski, E., Jongebloed, R., Tok, M., and Sikora, T., "Regularized Gradient Descent Training of Steered Mixture of Experts for Sparse Image Representation," in [*Accepted for publication in IEEE International Conference on Image Processing (ICIP '18)*], (2018).

[21] Viola, I., Rerabek, M., Bruylants, T., Schelkens, P., Pereira, F., and Ebrahimi, T., "Objective and subjective evaluation of light field image compression algorithms," in [*32nd Picture Coding Symposium*], (EPFL-CONF-221601) (2016).

[22] Rerabek, M. and Ebrahimi, T., "New light field image dataset," in [*8th International Conference on Quality of Multimedia Experience (QoMEX)*], (EPFL-CONF-218363) (2016).

[23] Verhack, R., Lange, L., Lambert, P., de Walle, R. V., and Sikora, T., "Lossless image compression based on kernel least mean squares," *2015 Picture Coding Symposium (PCS)* , 189–193 (2015).

[24] Wang, T.-C., Zhu, J.-Y., Kalantari, N. K., Efros, A. A., and Ramamoorthi, R., "Light field video capture using a learning-based hybrid imaging system," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)* **36**(4) (2017).

[25] Vetro, A., Wiegand, T., and Sullivan, G. J., "Overview of the stereo and multiview video coding extensions of the h.264/mpeg-4 avc standard," *Proceedings of the IEEE* **99**, 626–642 (2011).

[26] Wang, G., Xiang, W., Pickering, M., and Chen, C. W., "Light field multi-view video coding with two-directional parallel inter-view prediction," *IEEE Transactions on Image Processing* **25**, 5104–5117 (2016).

[27] Hansen, B. C. and Essock, E. A., "A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes.," *Journal of vision* **4 12**, 1044–60 (2004).