PROGRESSIVE MODELING OF STEERED MIXTURE-OF-EXPERTS FOR LIGHT FIELD VIDEO APPROXIMATION

Ruben Verhack*†, Glenn Van Wallendael*, Martijn Courteaux*, Peter Lambert*, and Thomas Sikora†
* Ghent University - imec, IDLab, Department of Electronics and Information Systems (ELIS), Belgium
† Technische Universität Berlin, Communication Systems Lab, Germany

***Abstract*—Steered Mixture-of-Experts (SMoE) is a novel framework for the approximation, coding, and description of image modalities. The future goal is to arrive at a representation for Six Degrees-of-Freedom (6DoF) image data. The goal of this paper is to introduce SMoE for 4D light field videos by including the temporal dimension. However, these videos contain vast amounts of samples due to the large number of views per frame. Previous work on static light field images mitigated the problem by hard subdividing the modeling problem. However, such a hard subdivision introduces visually disturbing block artifacts on moving objects in dynamic image data. We propose a novel modeling method that does not result in block artifacts while minimizing the computational complexity and which allows for a varying spread of kernels in the spatio-temporal domain. Experiments validate that we can progressively model light field videos with increasing objective quality up to 0.97 SSIM.**

I. Introduction

The user experience of virtual reality (VR) for camera captured content (e.g. 360° video) is currently limited, especially compared to the VR experience of computer generated scenes (e.g. in gaming). 360° video only enables three rotational head movements, compared to the desired Six Degrees-of-Freedom (6DoF), i.e. three translational movements (walking around) combined with three rotational movements (head movements). The 2D images observed by humans at each angle are processed versions of the higher-dimensional data the camera sensor has acquired. In terms of signal processing, we are likely presented with a high-dimensional sampling problem with nonuniform and nonlinear sample spacing and high-dimensional spatio-directionally varying sampling kernels [1]. The high-dimensional space is defined by the 5D plenoptic function [2]. However, when there are no occlusions (i.e. "open space" assumption), the 5D space can be reduced to the 4D light field [3], [4]. This assumption does not hold for 6DoF in large scenes, however, at the moment this is a widely used simplification [4].

Currently MPEG started standardization efforts for a 6DoF format [5]. Their envisioned process consists of two steps: (1) find the most important views on a scene, and (2) encode these views using well-known difference and transform coding approaches. At decoder side, views are synthesized potentially by using extra transmitted geometrical side-information. However, we argue that 2D regular sampling grids are not optimal representations for storing high-dimensional data, considering the observations mentioned above. Furthermore, we believe that the view synthesis process could place considerable computational complexity towards the decoder.

Therefore we previously introduced a novel methodology that aims to provide full 6DoF, namely *Steered Mixture-of-Experts* (SMoE). We directly model the underlying plenoptic function (or lower-dimensional projections of this function) in a continuous, analytical form [2]. We do so by identifying stationary regions of samples and by optimizing local linear regressors for that segment. The total regression corresponds to a smoothed piecewise linear approximation of the underlying function. Currently, we successfully applied SMoE on images, video, and 4D light field images [6]–[8]. In this paper, we introduce SMoE for light field video having a 5D coordinate space (time, two spatial, and two angular dimensions). As such we are nearing a full 6DoF representation.

Especially for light field images, SMoE was shown to yield competitive rate-distortion results for low- to mid-range bitrates [8]. Furthermore, the model takes on an informative structure due to the data-driven approach. It thus presents the decoder with MPEG-7 like descriptors including spatial edges, intensity flow, motion (video), and depth (light fields) [6]–[9]. For rendering it has three important properties [10]. Firstly, view-rendering is lightweight and pixel-parallel. Secondly, SMoE is a space-continuous representation, thus rendering at arbitrary resolution consists of merely sampling this function. Finally, all local light information at a certain point in the physical space is also localized in the model.

The encoding side consists of modeling the joint *probability density function* (pdf) of sample coordinates and sample amplitudes using a *Gaussian Mixture Model* (GMM). The estimated parameters of the model are then further compressed and binarized. The modeling phase is computationally challenging considering the enormous number of samples envisioned in 6DoF content. The light field videos handled in this paper consist of over 1 billion samples. Previous research on static content used hard subdivisions of the coordinate space to mitigate this problem [6], [8]. However, heavy block artifacts become visible when the scene is dynamic and other solutions should be considered.

In this paper, we present a novel progressive method to enable modeling of such large datasets using minibatches combined with local updates and split operations. As such, light

field videos are decomposed into high-dimensional kernels which hold spatially local light information in function of the viewing angle and the point in time.

II. Steered Mixture-of-Experts

A. Introduction

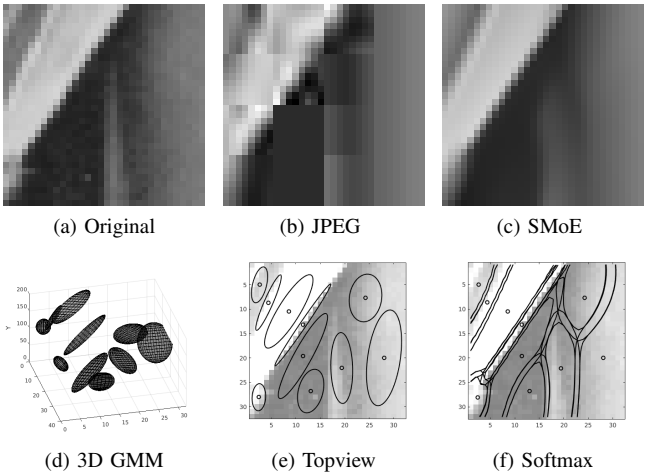
Steered Mixture-of-Experts (SMoE) is a novel framework for approximating image modalities with many applications, such as image modality coding, scale conversion (e.g. frame interpolation), and image description (e.g. depth estimation) [6]–[8]. Due to the sparse structure in SMoE, it is readily extendable 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, it 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.

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 [11]. Firstly, the input space is divided in soft-segments using a gating function. Secondly, local regressors (or *experts*) are sought that locally approximate the function optimally. Consequentially, the gating function lets experts collaborate in segments where they are trustworthy.

SMoE is based on the Bayesian, or "alternative" definition of the MoE model [11]. The Bayesian MoE approach models the joint probability of the input space X and the output space Y using a GMM. Each Gaussian kernel then simultaneously defines the gating function (soft-segmentation of X) and the local regressors (through the conditional probability function $Y|X$).

In SMoE, where the input space is the *coordinate space* (i.e. sample locations) and the output space is the *color space* (i.e. sample amplitudes), one such Gaussian then corresponds to one kernel as mentioned above. The gating function is thus defined by the probability of a coordinate to belong 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 sense.

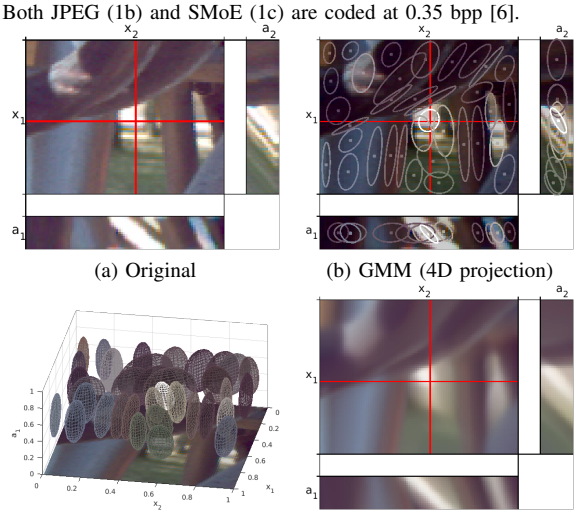
SMoE thus arrives at a sparse representation. The whole image modality is represented as a set of Gaussian kernels. These kernels are defined by their centers and their steering parameters. The coordinate space is 2D, 3D, or 4D in the case of respectively images, video, and static light fields [6]–[8], and analogously 5D for light field video. The color space for color images is conventionally represented as a 3D space, e.g. RGB or YCbCr. As the GMM models the joint probability



(a) Original (b) JPEG (c) SMoE

(d) 3D GMM (e) Topview (f) Softmax

Fig. 1. An example of the modeling with 10 components and reconstruction of a 32x32 pixel crop from *Lena* (1a). 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 (2c). Each kernel thus defines a 2D gradient as the *expert* function ($X \mapsto Y$). The gating function is defined by the soft-segmentation (1f). Both JPEG (1b) and SMoE (1c) are coded at 0.35 bpp [6].



(a) Original (b) GMM (4D projection)

(c) GMM (3D projection) (d) SMoE reconstruction

Fig. 2. SMoE applied on a spatial crop of a static 4D light field (*I01 Bikes* [12]), shown in an *epipolar* (EPI) representation in 2a. A GMM of the coordinates (4D) and color amplitudes (3D) is fit as shown in 2b using 35 kernels. A 3D reduction retaining the two spatial dimensions with only one angular dimension shown in 2c. Finally, the regression based on the GMM is shown in 2d. Note how the white background is approximated by a single kernel [13].

of the 5D coordinate and 3D color space, we thus arrive at 8D Gaussian kernels. The parameters of these kernels are typically estimated using computational efficient variations of the *Expectation-Maximization* (EM) algorithm [14]. Due to this likelihood optimization, kernels will steer along the dimensions of the highest correlation, e.g., along spatial or temporal consistencies.

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 [6]. Fig. 2 illustrates SMoE applied to static 4D light fields [8].

B. Theory

The goal of regression is to optimally predict a dependent random vector $Y \in \mathbb{R}^q$ from a known random vector $X \in \mathbb{R}^p$. In SMoE, X corresponds to pixel coordinates (i.e., the 5D coordinate space) and Y to the pixel amplitudes (i.e., the 3D color space). The joint probability function of the coordinate space X and color space Y is modeled as a multi-modal, multi-variate GMM. Each Gaussian kernel then defines a soft-segment in X and a local regressor ($X \mapsto Y$). The local regressor is defined by a measure of central tendency (e.g. the mean, median, mode) of the conditional pdf $Y|X$. In this paper, we will limit the case to the mean-estimator, i.e. $\mathbf{E}[Y|X = \boldsymbol{x}]$.

Let us assume $D = \{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=1}^N$ to be N pixels to be modeled with coordinates $\boldsymbol{x}$ and amplitudes $\boldsymbol{y}$:

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

$$\text{and } \sum_{j=1}^{K} \pi_j = 1, \boldsymbol{\mu}_j = \begin{bmatrix} \boldsymbol{\mu}_{X_j} \\ \boldsymbol{\mu}_{Y_j} \end{bmatrix}, R_j = \begin{bmatrix} R_{X_jX_j} & R_{X_jY_j} \\ R_{Y_jX_j} & R_{Y_jY_j} \end{bmatrix}$$

The parameters of this mixture model with K Gaussian distributions are $\Theta = [\theta_1, \cdots, \theta_K]$, with $\theta_j = (\pi_j, \boldsymbol{\mu}_j, R_j)$, being the population densities, centers, and covariances respectively.

The conditional pdf of the mixture model $Y|X$ is used to derive the regression function [15], [16]:

$$p_Y(\boldsymbol{y}|X = \boldsymbol{x}) = \sum_{j=1}^{K} w_j(\boldsymbol{x}) \mathcal{N}(\boldsymbol{y}; m_j(\boldsymbol{x}), \tilde{R}_{Y_j,Y_j}) \tag{2}$$

with mixing weights $w_j(\boldsymbol{x})$, regressors $m_j(\boldsymbol{x})$, and conditional covariance $\tilde{R}_{Y_j,Y_j}$:

$$w_j(\boldsymbol{x}) = \frac{\pi_j \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{X_j}, R_{X_jX_j})}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{X_i}, R_{X_iX_i})} \tag{3}$$

$$m_j(\boldsymbol{x}) = \boldsymbol{\mu}_{Y_j} + R_{Y_jX_j} R_{X_jX_j}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{X_j}), \tag{4}$$

$$\tilde{R}_{Y_j,Y_j} = R_{Y_jY_j} - R_{Y_jX_j} R_{X_jX_j}^{-1} R_{X_jY_j} \tag{5}$$

The regression of the model is defined as the expected value y given a sample location x through the conditional. From Eq. 2 and 3 follows the regression function $m(\boldsymbol{x})$:

$$\hat{\boldsymbol{y}} = m(\boldsymbol{x}) = \mathbf{E}[Y|X = \boldsymbol{x}] = \sum_{j=1}^{K} w_j(\boldsymbol{x}) m_j(\boldsymbol{x}) \tag{6}$$

A signal at location x can be predicted by the weighted sum over all K mixture components (Eq. 6). Every component in the mixture model is considered as an expert and the experts collaborate towards the definition of the regression function.

III. Progressive Modeling

SMoE models are typically trained using variants of the EM-algorithm [14]. Previous research for modeling light fields was based on a hard subdivision of the coordinate space into B independent blocks, and thus consisted of a number smaller independent modeling tasks [8]. Firstly, this subdivision drastically lowers the computational complexity. The likelihood is calculated between each sample in the block N_{block} and each kernel in that block K_{block} in each iteration. This results in $O(K_{\text{b}} N_{\text{b}})$ evaluations per block, instead of $O(KN)$ with $N = \sum_{m=1}^{B} N_{b_m}$ and $K = \sum_{m=1}^{B} K_{b_m}$. Secondly, the subdivision allows to spent varying kernel "budgets" on blocks depending on the spatial variance in that block [6]. However for modalities with a time dimension, the block-division results in disturbing block artifacts on moving objects (Fig. 3).

TABLE I
Properties of SMoE modeling techniques

	global[7]	**block-wise**[6], [8]	**proposed**
Varying density of kernels	low	high	high
Complexity	high	low	low
Block artifacts	no	yes	no

In this work, we implement the training using minibatches instead of batch updates. Previous research has shown that updating the kernel parameters based on a subset of the samples heavily increases the robustness, while drastically lowering the computational demands [17]–[19]. The minibatch approach introduces two hyperparameters (s, α): the minibatch-size s (with $s \ll N$) and α which drives the learning speed.

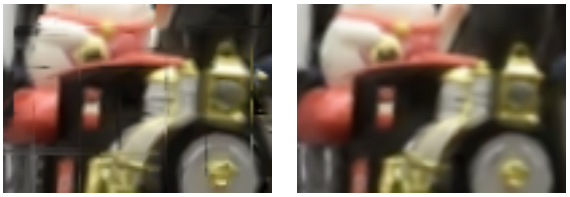
We propose in the next subsections a computational efficient and global EM variant for training SMoE models that also allows for an varying distribution of the kernels. Firstly, we simulate global modeling by performing block-wise updates. Secondly, we implement an iterative train-and-split strategy in order to achieve an varying distribution of the kernels. Table I summarizes the features of the discussed techniques.

A. Block-level updates

The following optimization allows us to drastically lower the computational demands for one minibatch iteration. The light field video is subdivided in overlapping spatio-temporal blocks, e.g. 32x32 pixels over 32 frames. These blocks are visited consecutively. A minibatch sample is selected from this block. The loglikelihood of each sample is determined by evaluating only the nearby relevant kernels. The loglikelihood of other kernels with these samples is considered to be zero. The relevant kernels are the kernels that have a center within a spatio-temporal relevance window. The kernels in the relevance window are updated after each block visit. As such, only the set of relevant kernels K_b and the s local minibatch samples are needed in memory. This results in $O(K_b s)$ per iteration per block, which heavily reduces the original requirements of $O(KN)$ per iteration. Note that kernels can migrate over the whole domain and can be present in several relevance windows in one image pass.

B. Kernel splitting

For the application of image approximation, it is desired to minimize the prediction variance of $Y|X = \boldsymbol{x}$. Previous research suggested the beneficial varying kernel spread property of initializing the EM algorithms using a split approach [20]. Borrowing ideas from these observations, we developed



(a) block-based (b) proposed

Fig. 3. The hard block-level subdivision of the coordinate space results in visually disturbing artifacts when objects cross block boundaries in time. This is mitigated by the block-level updates that simulate global modeling.

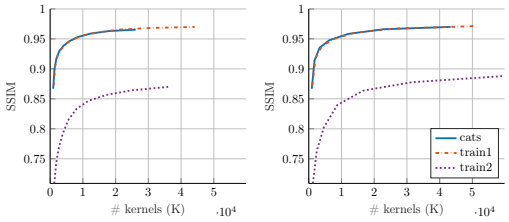


Fig. 4. This figure illustrates the SSIM results for the three light field videos modeled using progressive modeling introduced in Sec. III. The modeling was done by splitting respectively 10% (left) and 30% (right) of the top uncertain kernels.

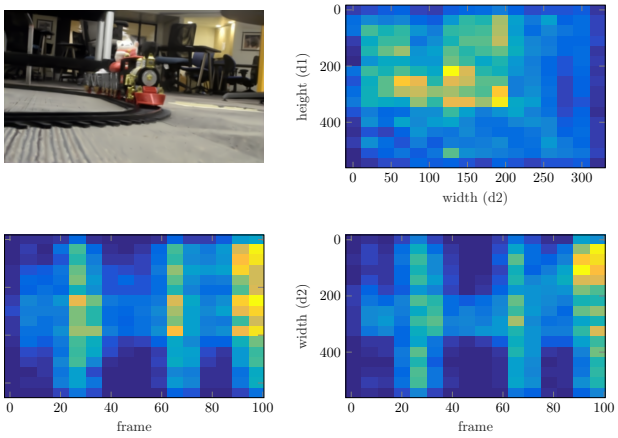


Fig. 5. This figure illustrates the density of the kernels along the spatial and time dimension for a model trained on *train2* (top left) using $K = 49,455$ kernels. The density of kernels measures the number of kernels in a given image area. The light field depicts a toy train coming from the background center towards the front left. It is clear that the density is greater in areas with high motion due to the split operations. Spatially the kernels are concentrated top left where the train rides. Over time the train comes closer to the camera, this results in more kernels are spent on the later frames. The train's trajectory is even visible on the bottom right density map. The train is first in the center and then moves towards the left (lower d_2 value).

a progressive modeling strategy that progressively creates models with increasingly higher number of kernels where the prediction variance is high. This is done by splitting a certain amount of kernels based on the weighted variance of the luma-channel of their conditional variance $\pi_j \tilde{R}^{(1,1)}_{Y_j,Y_j}$ (Eq. 5). Note that this calculation is significantly cheaper than the calculating the prediction error as in [20].

To conclude, we let the modeling start with an initial number of kernels K_{init} and model using block-level updates using minibatches. After convergence, we split the most uncertain kernels into four kernels of 1/4th of the original size. These kernels are displaced from the original center along the time and spatial dimensions based on the variance of the original kernels in those dimensions. These kernels then serve as a new initialization. This is process is repeated until the number of kernels reaches a predetermined K_{max}.

IV. Experimental Evaluation

A. Dataset

We selected three light field videos *cats*, *train1*, *train2* [21]. These light field videos are reconstructed from Lytro light field images taken at 3fps, combined with a DSLR capturing a view at 30 fps. We use these videos as the ground truth in this experimental section, even though they contain some artifacts from the view synthesis process. These videos have 8x8 views, between 80 and 110 frames, and a spatial resolution around 544x320 pixels. Each light field video thus contains roughly 1 billion samples.

B. Experiments

Fig. 4 shows the results for the three light fields being modeled progressively using the approach introduced in Sec. III. All models were initialized using $K_{\text{init}} = 2048$ and trained using $(s, \alpha) = (1e4, .6)$. Two splitting ratios were used: 10% and 30%. Due to the high number of views ($\pm$100x8x8), we measured the average SSIM for a single view in each frame, rotating over the views $(2,2), (3,6), (4,4), (7,2)$.

The block-level updates were done using spatio-temporal blocks of 36x36x36 pixels with an overlap of four pixels in each dimension. The kernel relevance window that determines which kernels to involve in this update was set to 54x54x54 pixels. It is clear that subsequent models introduce a steady increase of reconstruction quality up to 0.97 SSIM. Fig. 4 also suggests that for this setup, the split-ratio is less important. As such, larger split-operations (right), which require less meta-iterations, do not seem to compromise the quality.

Fig. 5 illustrates the kernel distribution over the spatial and time dimensions for one model of *train2*. The vertical lines of high density of kernels along the time dimensions on the bottom row result from the block-level updates. In static areas kernels spread as far as possible in the time dimension as there is no change in color intensities. However, due to kernels falling outside of the relevance window, kernels in static regions have a limited spread and other kernels take over. The position of the kernel centers on the angular dimensions are located near the middle, suggesting that each kernel has maximal stretch along the angular dimensions, similar to Fig. 2.

V. Conclusion

In this paper, we have shown that SMoE is extendable to light field videos when using a smart local updating strategy and by progressively incrementing the number of kernels in regions with higher spatial and temporal changes. Experiments have shown that each kernel specializes on the angular light information in one particular spatio-temporal region of varying size. Furthermore, the temporal variances stretch maximally as desired for kernels in static image regions.

REFERENCES

[1] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: briefly revisiting 25 years of research," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 59–69, 2016.

[2] E. Adelson and J. Bergen, *The plenoptic function and the elements of early vision*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991.

[3] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96*, New York, New York, USA: ACM Press, 1996, pp. 31–42.

[4] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: an overview," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2017.

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

[6] R. Verhack, T. Sikora, L. Lange, G. Van Wallendael, and P. Lambert, "A universal image coding approach using sparse steered mixture-of-experts regression," in *IEEE Proc. Int. Conf. on Image Processing (ICIP)*, IEEE, 2016, pp. 2142–2146.

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

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

[9] T. Sikora, "The mpeg-7 visual standard for content description – an overview," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 696–702, 2001.

[10] V. Avramelos, R. Verhack, I. Saenen, G. Van Wallendael, B. Goossens, and P. Lambert, "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] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 23, no. 8, pp. 1177–1193, 2012.

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

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

[14] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[15] H Sung, "Gaussian mixture regression and classification," PhD thesis, Rice University, 2004.

[16] G. Bugmann, "Normalized gaussian radial basis function networks," *Neurocomputing*, vol. 20, no. 1-3, pp. 97–110, 1998.

[17] R. Verhack, T. Sikora, L. Lange, G. Van Wallendael, and P. Lambert, "Steered mixture-of-experts for 4-d light field approximation, coding, and description," *Submitted to Transactions on Multimedia*, 2018.

[18] P. Liang and D. Klein, "Online em for unsupervised models," in *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, Association for Computational Linguistics, 2009, pp. 611–619.

[19] M.-a. Sato and S. Ishii, "On-line em algorithm for the normalized gaussian network," *Neural Computation*, vol. 12, no. 2, pp. 407–432, 2000.

[20] R. Jongebloed, R. Verhack, L. Lange, and T. Sikora, "Hierarchical learning of sparse image representations using steered mixture-of-experts," in *Submitted to International Conference on Multimedia and Expo '18*, 2018.

[21] T.-C. Wang, J.-Y. Zhu, N. K. Kalantari, A. A. Efros, and R. Ramamoorthi, "Light field video capture using a learning-based hybrid imaging system," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.