AN MSE APPROACH FOR TRAINING AND CODING STEERED MIXTURES OF EXPERTS

Michael Tok, Rolf Jongebloed, Lieven Lange, Erik Bochinski, and Thomas Sikora
Communication Systems Group
Technische Universität Berlin
Berlin, Germany

***Abstract*—Previous research has shown the interesting properties and potential of Steered Mixtures-of-Experts (SMoE) for image representation, approximation, and compression based on EM optimization. In this paper we introduce an MSE optimization method based on Gradient Descent for training SMoEs. This allows improved optimization towards PSNR and SSIM and de-coupling of experts and gates. In consequence we can now generate very high quality SMoE models with significantly reduced model complexity compared to previous work and much improved edge representations. Based on this strategy a block-based image coder was developed using Mixture-of-Experts that uses very simple experts with very few model parameters. Experimental evaluations shows that a significant compression gain can be achieved compared to JPEG for low bit rates.**

***Index Terms*—Image Compression, Steered Mixture of Experts, Image Regression, Machine Learning**

I. Introduction

Over the last 30 years still image coding including compression of multiple-amplitude level images has been researched intensively. Today the JPEG standard [12] is still the most widely deployed compression approach for coding natural imagery. The more recent JPEG2000 [9] standard compression scheme improves compression efficiency over JPEG. Both standards are based on the well-known block-based "frequency domain" transform coding philosophy that today dominates research in the field of image and video compression [7].

Our challenge is the research into novel 2D as well as 3D and N-D image and video representations that may pave the way towards more efficient next generation compression strategies. To this end we focus on non-linear sparse representations of imagery using machine learning algorithms for optimization.

In this context we have recently introduced the so-called "Steered-Mixture-of-Experts" (SMoE) framework for sparse modeling, regression and coding of imagery. Steered experts can "steer" along edges in N-D imagery and provide excellent edge-aware image reconstruction properties. When used for compression the SMoE model parameters are quantized and coded and used directly for reconstruction at the decoder. As such, the compression approach departs drastically from existing JPEG-like "transform"-based coding approaches. The SMoE models explain the data in the spatial rather than in the transform domain. An interesting feature of the steering capabilities of the SMoE kernels is their descriptive nature about correlation in the imagery. When used for coding they provide MPEG-7-like low- and mid-level image descriptors on bit-level at the decoder [6] [8].

Our previous work on SMoEs applied to image compression [11], video compression [5] and light-field coding [10] showed that SMoE compression can achieve significant gains compared to existing transform-based approaches.

Steered Mixtures-of-Experts can be seen as stochastic neural networks that are comprised of so-called "gating functions" and associated "steering experts" [4]. In our approach applied to imagery we employ Gaussian kernels as experts to explain the image data in the 2D (or N-D) regions defined by the associated (Gaussian kernel) gating functions. The SMoE approach follows the divide-and-conquer principle. All experts and gating functions collaborate towards reconstruction of the image data. Given a fixed number of experts allocated for reconstruction of a 2D or N-D image, the parameters of the network (location, the steering parameters and weights of experts in N-D space) need to be identified. The well-known Expectation-Maximization (EM) algorithm is usually employed to optimize the experts parameters [3].

In our previous work N-D Gaussian kernel functions were used to jointly represent the experts and associated gates. With this assumption the SMoE modeling approach is equivalent to a Gaussian-Mixture-Model (GMM) with steered Gaussians used for regression. Gates and experts are strictly coupled in this representation. However, the estimation of the steering model parameters using the well-known Expectation Maximization algorithm is not necessarily optimal when the approach is used for compression of signals, as it maximizes the likelihood function rather than minimizing the Mean Squared Error (MSE). Also the strict coupling of "experts" and "gating functions" previously used imposes unnecessary restrictions for modeling and coding.

In this paper we introduce an MSE optimization method based on Gradient Descent for training SMoEs. This allows improved optimization towards PSNR and SSIM and de-coupling of experts and gates. In consequence we can now depart from the traditional GMM model and generate very high quality SMoE models with significantly reduced model complexity compared to previous work and much improved edge representations. Based on this strategy a block-based image coder is developed using Mixture-of-Experts that uses very simple experts with very few model parameters.

II. Theoretical Background

Gaussian Mixture Models can be used to derive Steered-Mixtures-of-Experts, i.e. for gray level images. Here GMMs define a multivariate joint density distribution for (spatial input) random vector $\boldsymbol{x}$ and (luminance output) random variable y as sum of K weighted 3D Gaussian distributions (our desired 3D steered experts based on Gaussian kernel functions)

$$p\left(\boldsymbol{x}, y\right) = \sum_{i=1}^{K} \pi_i \cdot \mathcal{N}\left(\boldsymbol{x}, y; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \tag{1}$$

with mixing coefficients π_i, for which $\sum_{\forall i} \pi_i = 1$, covariance matrices

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_{XX,i} & \boldsymbol{\Sigma}_{XY,i} \\ \boldsymbol{\Sigma}_{XY,i}^T & \sigma_{Y,i}^2 \end{bmatrix} \tag{2}$$

and mean values (centers)

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_{X,i} \\ \mu_{Y,i} \end{bmatrix}. \tag{3}$$

When the model parameters are trained (i.e. using EM algorithm) a 2D regression function can be determined. i.e. using the expected value of the conditional distribution of Y given $\boldsymbol{X}$

$$y_p\left(\boldsymbol{x}\right) = \mathbb{E}\left[Y|\boldsymbol{X}\right] = \sum_{i=1}^{K} m_i\left(\boldsymbol{x}\right) \cdot w_i\left(\boldsymbol{x}\right) \tag{4}$$

with the so called gradient or (hyper-) plane components

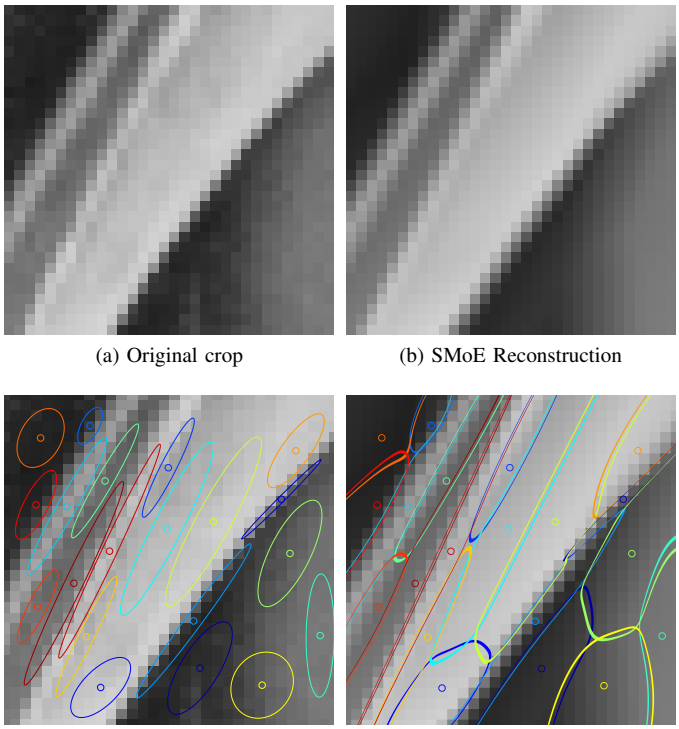
$$m_i\left(\boldsymbol{x}\right) = \mu_{Y,i} + \boldsymbol{\Sigma}_{XY,i}^T \boldsymbol{\Sigma}_{XX,i}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_{X,i}\right) \tag{5}$$

and weighted soft max gating functions

$$w_i\left(\boldsymbol{x}\right) = \frac{\pi_i \cdot \mathcal{N}\left(\boldsymbol{x}, y; \boldsymbol{\Sigma}_{XX,i}, \boldsymbol{\mu}_{X,i}\right)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}\left(\boldsymbol{x}, y; \boldsymbol{\Sigma}_{XX,j}, \boldsymbol{\mu}_{X,j}\right)}. \tag{6}$$

Notice that experts and gates are derived from the 3D kernel functions and they are thus coupled using such GMMs. Figure 1 depicts as an example the reconstruction of a 32×32 pixel block of *Lena* with $K = 20$ steering kernels trained using the EM algorithm. The 3D steering kernels steer also into y-direction but in the figures only projections onto the $\boldsymbol{x}$-plane are shown. It is apparent that the components steer along direction of highest correlation, in particular in the neighbourhood of edges. Even though softmaxed gating functions are derived from the same Gaussian kernels they define arbitrarily-shaped windows. In non-overlapping areas between neighbouring gates sharp edges can be reconstructed. With overlapping gates smooth transition between segments are provided, as in the lower right area of the image.

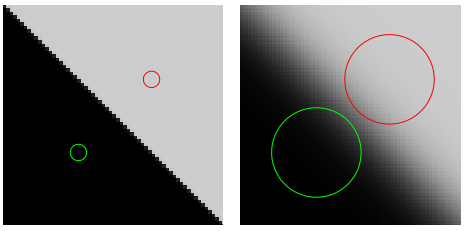
Figure 2 depicts that even very simple radial experts can reproduce edges in images. Two radial kernels with three parameters each (plus one bandwidth parameter for both) can easily reproduce directional edge patterns with sharp (Fig. 2(a)) or smooth (Fig. 2(b)) transitions without ringing artifacts known from JPEG and JPEG 2000. Even though



(a) Original crop (b) SMoE Reconstruction

(c) Position and steering of kernels (d) Soft-maxed gating functions

Fig. 1: SMoE Modeling example with $K = 20$ components



(a) Small bandwidth (b) Large bandwidth

Fig. 2: Edge reconstruction with simple radial kernels

the kernels are radial the gates steer along the edges. This attractive property of radial kernels is explored further below for modeling and coding.

A. Separating Experts and Gates

The original Mixture-of-Experts approach invented by Jacobs and Jordan is a very flexible framework applied to classification, regression and prediction of signals [13]. In general it is possible to build SMoE models by defining separate experts and gates, $y_p\left(\boldsymbol{x}\right) = \sum_{i=1}^{K} m_i\left(\boldsymbol{x}\right) \cdot w_i\left(\boldsymbol{x}\right)$, with simple experts and powerful but expensive gates or vice versa. It is not even necessary that gates and experts are derived based on a statistical framework using density functions, such as with GMM in the previous section. Notice that in this case the model parameters may not be trainable using EM.

In the two dimensional case (as for images), equation 5 describes a plane in the form

$$m_i(\boldsymbol{x}) = m_{0,i} + m_{1,i} \cdot x_1 + m_{2,i} \cdot x_2 \tag{7}$$

with coefficients

$$m_{0,i} = \mu_{Y,i} - \boldsymbol{\Sigma}_{XY,i}^T \boldsymbol{\Sigma}_{XX,i}^{-1} \boldsymbol{\mu}_{X,i} \tag{8}$$

$$\begin{bmatrix} m_{1,i} & m_{2,i} \end{bmatrix} = \boldsymbol{\Sigma}_{XY,i}^T \boldsymbol{\Sigma}_{XX,i}^{-1} \tag{9}$$

This couples the gating function and expert by shared parameters (see eq. 6).

Separated experts can vary from simple, constant offsets

$$m_c = m_0 \tag{10}$$

to polynomial models e.g. with a degree of 2:

$$m_p = m_0 + m_1 x_1 + m_2 x_2 + m_3 x_1 x_2 + m_4 x_1^2 + m_5 x_2^2 \tag{11}$$

or functions with even higher complexity. In the same manner, the gating can be varied and built even on neural network sigmoid functions. A straight-forward approach used in this paper is to use radial kernels to generate gates with covariance matrices

$$\boldsymbol{\Sigma}_{XX,i} = \sigma_i^2 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{12}$$

inducing gating functions from radial kernels as in Figure 2.

B. MSE Optimization of SMoEs

In this paper we introduce an MSE optimization method based on Gradient Descent for training SMoEs. Our purpose is twofold. Firstly, we like to be able to train SMoEs with separated experts and gates that cannot be optimized using EM. Secondly, in a regression framework for compression, it seems advisable to use an optimization criteria that is better tuned towards PSNR and SSIM criteria to reconstruct images with improved quality compared to EM algorithm with likelihood criteria. We define the objective function $\mathcal{L}$ as

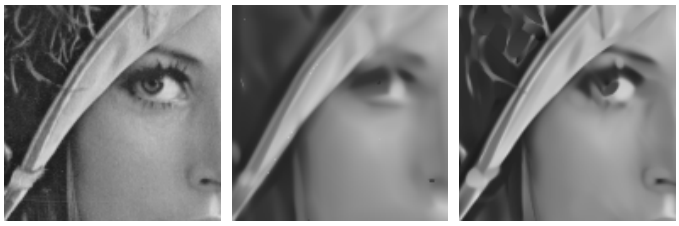
$$\mathcal{L} := \frac{1}{N} \sum_{n=1}^{N} \left(y_n - \sum_{i=1}^{K} m_i(\boldsymbol{x}) \cdot w_i(\boldsymbol{x}_n) \right)^2 . \tag{13}$$

To find a set of parameters π_i, $\boldsymbol{\Sigma}_i$, $\boldsymbol{m}_i$ and $\boldsymbol{\mu}_i$ that minimize

$$\underset{\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{m},\boldsymbol{\pi}}{\arg\min} \{\mathcal{L}\} \tag{14}$$

we employ the negative gradient $-\nabla\mathcal{L}$ via a gradient descent approach for this purpose.

We conducted extensive tests to understand the performance of the MSE training algorithm compared to EM. In a first set of experiments we were interested in comparing full statistical SMoE GMM models (9 parameters for each Gaussian). For this purpose the proposed MSE based optimization as well as the methods of [1] and [2] were used. Figure 3 provides a representative snapshot of results on a 128×128 pixel crop of grayscale Lena in Figure 3(a). We used 100 components



(a) Original crop (b) GMM-EM, 100 components (25.59 dB) (c) GMM-MSE, 100 components (31.83 dB)

Fig. 3: Comparison of SMoE reconstruction qualities for EM and for MSE optimization with $K = 100$ components

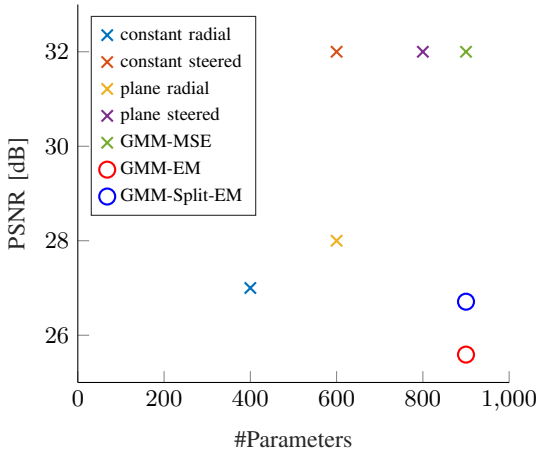


Fig. 4: Reconstruction quality overview for various expert-gate-combinations ($K = 100$) (o - learned with EM [1], o - learned with [2])

for each of the models with their centers $\boldsymbol{\mu}_i$ initialized on an evenly distributed grid. Figure 3(b) depicts the reconstruction with a model trained using EM [1] (GMM-EM) and in Figure 3(c) our result (GMM-MSE) using MSE Gradient Descent. The traditional EM approach is outperformed by more than 6 dB. In particular the edges are reconstructed with remarkably improved sharpness.

These models are again depicted in Figure 4 (in total 900 parameters each) for same image with the attempt to compare those results with models of less complexity. The GMM-Split-EM algorithm [2] improves over GMM-EM by 1dB with same GMM model. On the other hand it turns out that separating experts from gates indeed provides benefits. In comparison to the full model GMM-MSE result, a MSE-trained model without GMM weights (8 parameters, plane experts and steered gates) and an even less complex model without weights (6 parameters, constant experts, steered gates) achieve essentially identical performance. For coding, models with less parameters are favourable. Here the most simple MSE model with 4 parameters (constant experts, radial gates) turns out to be of particular interest, since it combines reasonable performance with very low complexity. This model is fast to train and coded with few parameters.

III. Image Compression using Radial SMoE Models

Our purpose is to use MSE trained SMoEs for coding of imagery. As observed in the previous sections, a pair of very simple radial kernels in SMoE can already reproduce sharp edges (Figure 2) as they often appear in natural images. These simple models require for coding 3 parameters (location and grey level value (eq. 10) as well as one shared parameter for the kernel bandwidth (eq. 12). Results in Figure 4 indicate that these MSE models provide reasonable performance.

In this section we introduce an adaptive block-based SMoE image coder that employs these simplest kernels. Each grey level image is divided into non-overlapping blocks. In the encoder we calculate the variance of the amplitudes in each block. If the variance is below a given threshold the blocks are assigned to a background non-textured class. Otherwise these blocks are textured blocks.

For each textured block, a set of K constant experts with radial gates are trained:

$$y_p(\boldsymbol{x}) = \sum_{i=1}^{K} m_i \cdot \frac{\exp\left(-S\left\|\boldsymbol{x}-\boldsymbol{\mu}_i\right\|^2\right)}{\sum\limits_{j=1}^{K}\exp\left(-S\left\|\boldsymbol{x}-\boldsymbol{\mu}_j\right\|^2\right)}. \tag{15}$$

For optimization the gradients for gate center position and expert optimization have the form:

$$\nabla_{\boldsymbol{\mu}_k}\mathcal{L} = 4S\sum_{n=1}^{N} e_n \cdot w_{n,k} \cdot (y_p(\mathbf{x}_n) - m_k(\mathbf{x}_n)) \cdot (\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{16}$$

$$\nabla_{m_k}\mathcal{L} = -2\sum_{n=1}^{N} e_n \cdot w_{n,k} \tag{17}$$

with the regression error $e_n = (y_n - y_p(\boldsymbol{x_n}))$ and weight coefficients

$$w_{n,k} = \frac{\exp\left(-S\left\|\mathbf{x}_n-\boldsymbol{\mu}_k\right\|^2\right)}{\sum_{j=1}^{K}\exp\left(-S\left\|\mathbf{x}_n-\boldsymbol{\mu}_j\right\|^2\right)}. \tag{18}$$

For non-textured blocks only the average block luminance value is encoded. For the textured blocks the center amplitudes as well as the locations are quantized and coded after optimization. Besides a small amount of header information, the final compressed bitstream only contains flags to distinguish between textured and non-textured blocks, mean luminance values for non-textured blocks and quantized model parameters for textured blocks. All model parameters are PCM coded, thus without exploring redundancies between parameters.

IV. Experimental Evaluation

Four well known test images with a resolution of 512×512 pixels have been selected to illustrate the encoding performance of the proposed MoE based image compression framework[1]. The block size is set to 16×16 and each textured block is modeled by four experts. Each center position component and each constant expert is encoded with 3 to 5 bits. The texture standard deviation threshold varies between 4 and 20. All kernels share the same bandwidth $S = 0.018$.

[1]For supplemental material and sample code please visit http://www.nue.tu-berlin.de/research/gmm_image_compression/

TABLE I: Bit Rate and Quality examples

	JPEG			Proposed		
Sequence	Rate [bpp]	PSNR [dB]	SSIM	Rate [bpp]	PSNR [dB]	SSIM
Baboon	0.14	21.31	0.45	0.14	22.55	0.49
	0.16	21.76	0.50	0.16	22.64	0.50
Cameraman	0.15	26.45	0.73	0.08	26.47	0.80
	0.17	27.88	0.81	0.09	26.30	0.81
Lena	0.14	24.82	0.67	0.14	27.95	0.77
	0.17	27.32	0.74	0.17	28.11	0.78
Peppers	0.14	24.95	0.62	0.14	28.29	0.71
	0.17	27.46	0.69	0.17	28.46	0.75

Figure 5 shows the R-D performance of our MoE based image coding framework in comparison to JPEG in terms of PSNR and SSIM for the test images *Peppers* and *Cameraman*. Table I compares a selection of R-D points. In lower bit rate ranges the proposed framework significantly outperforms JPEG both in terms of PSNR and SSIM. This is remarkable, because no redundancy coding has been implemented as yet. For higher bit rate ranges the model with only four kernels per textured block is insufficient to reproduce fine textures.

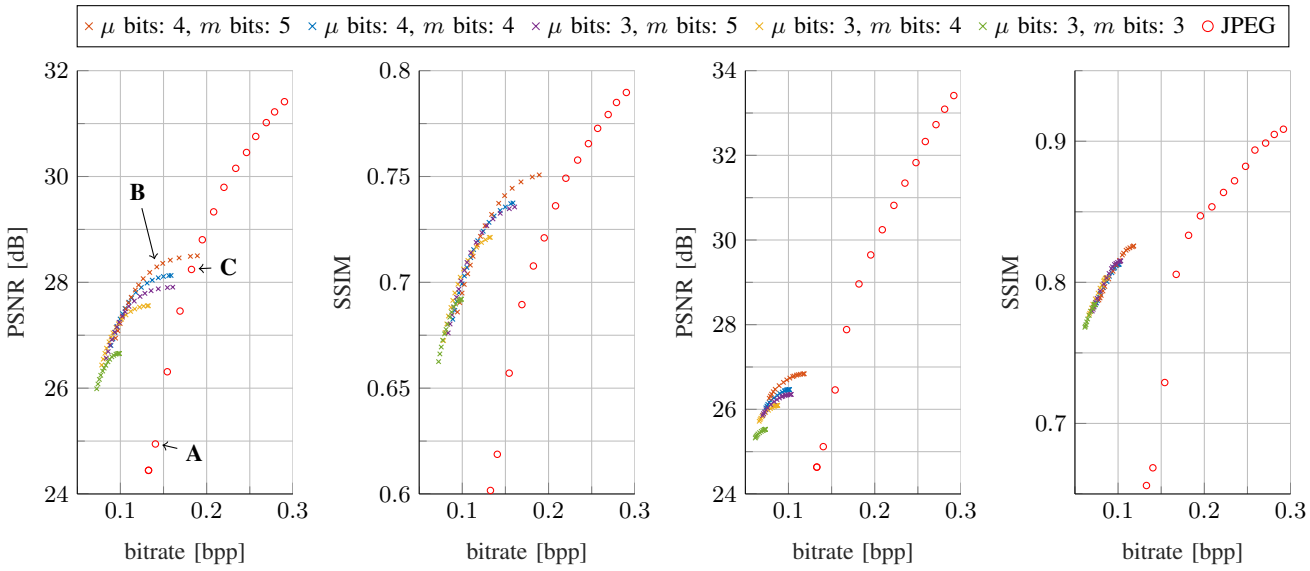
Figure 6 provides visual comparison and confirms that at lower rates the quality using MSE SMoE is significantly improved compared to JPEG. In particular the clear and crisp edges without "JPEG/JPEG2000 ringing" reproduced by the four very simple SMoE kernels in each 16x16 textured block are impressive. The corresponding rate distortion points of these visual examples are marked as (**A**), (**B**) and (**C**) in Figure 5(a).

V. Summary and Conclusion

A novel MSE training method for SMoE modelling and compression was presented in this paper. Results show that images can be reconstructed with drastically improved quality compared to the standard EM algorithm. In addition this allows for training of SMoE models with separated experts and gates. When implemented into an image coder the simplest experts and associated gates provide excellent image quality with very sharp edges at low rates. The image coding approach presented fails to perform well at higher rates. In future work we plan to make the coder efficient by adaptive assignment of varying numbers of kernels (thus bits) to blocks with different texture details. We also envision to adaptively incorporate more complex steering experts to the system.

VI. Acknowledgements

This work was supported by a Google Faculty Research Award 2016 in Machine Perception.

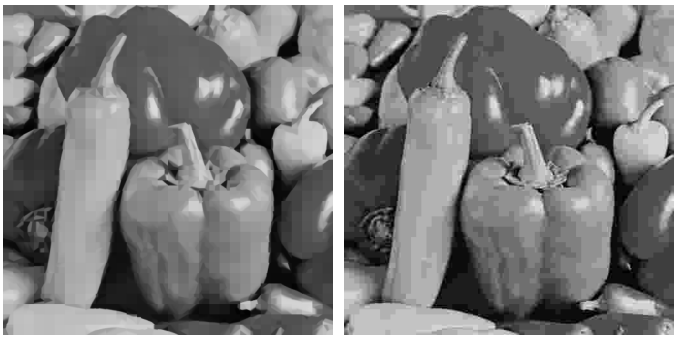


(a) PSNR Peppers (b) SSIM Peppers (c) PSNR Cameraman (d) SSIM Cameraman

Fig. 5: Rate distortion curves (PSNR and SSIM) for two selected test images

(a) Peppers Original

(b) Peppers JPEG at 0.14 bpp, 24.95 dB, SSIM: 0.62 (**A**)



(c) Peppers MoE at 0.14 bpp, 28.29 dB, SSIM: 0.74 (**B**)

(d) Peppers JPEG at 0.18 bpp, 28.24 dB, SSIM: 0.71 (**C**)

Fig. 6: Original (a) and reconstruction examples of Peppers encoded with JPEG (b), (d) and the proposed MoE based image coding framework (c)

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[2] R. Jongebloed, R. Verhack, L. Lange, and T. Sikora. Hierarchical Learning of Sparse Image Representations using Steered Mixture-of-Experts. In *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2018.

[3] M. Jordan and L. Xu. Convergence Results for the EM Approach to Mixtures of Experts Architectures. *Neural Networks*, 8(9):1409–1431, jan 1995.

[4] M. I. Jordan and R. A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6(2):181–214, 1994.

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

[6] T. Sikora. The MPEG-7 Visual Standard for Content Description - An Overview. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):696–702, 2001.

[7] T. Sikora. Trends and perspectives in image and video coding. *Proceedings of the IEEE*, 93(1):6–17, Jan 2005.

[8] T. Sikora and L. Chiariglione. MPEG-4 video and its potential for future multimedia services. In *Circuits and Systems, 1997. ISCAS '97., Proceedings of 1997 IEEE International Symposium on*, volume 2, pages 1468–1471 vol.2, Jun 1997.

[9] A. Skodras, C. Christopoulos, and T. Ebrahimi. The jpeg 2000 still image compression standard. *IEEE Signal Processing Magazine*, 18(5):36–58, Sep 2001.

[10] R. Verhack, T. Sikora, L. Lange, R. Jongebloed, G. V. 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)*, pages 1183–1188, July 2017.

[11] R. Verhack, T. Sikora, L. Lange, G. V. Wallendael, and P. Lambert. A universal image coding approach using sparse steered Mixture-of-Experts regression. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2142–2146, Sept 2016.

[12] G. K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, Feb 1992.

[13] S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, Aug 2012.