Dictionary Learning based High Frequency Inter-Layer prediction for Scalable HEVC

Jens Schneider, Johannes Sauer, Mathias Wien

Institut für Nachrichtentechnik
RWTH Aachen University, Germany

schneider@ient.rwth-aachen.de

***Abstract*—Image scale-up is a crucial task in resolution varying scalable video coding, as the coding costs for the enhancement layer depend heavily on the prediction signal generated by inter-layer prediction. In order to generate a suitable prediction signal the missing high frequencies in the base layer picture have to be reconstructed. For this purpose upscaling methods which go beyond the classical sampling theory are required. In this paper, an image scale-up method based on dictionary learning and sparse coding techniques for inter-layer prediction in scalable video coding is presented. Experimental results show that the proposed method outperforms state of the art scalable coding models in the case of 2x upscaling. In more detail, $2.35\,\%$ BD-rate savings against SHM 12.0 reference software are observed on average for an All Intra coding configuration. The maximum achieved rate savings were $6.15\,\%$ for the sequence PeopleOnStreet.**

***Index Terms*—inter-layer prediction, high frequency prediction, scalable video coding, dictionary learning, super-resolution**

I. Introduction

Scalable video coding aims at coding a video signal at different resolutions into one bitstream, such that parts of the bitstream can be discarded but the stream is still decodable. In order to achieve an efficient representation of the video at different scales, inter-layer prediction is applied. For this purpose, image scale-up methods reducing the residual for the enhancement layer as much as possible are required. In the scalable extension of the High Efficiency Video Coding (HEVC) standard [1], interpolation filters are used for inter-layer prediction [2]. Those interpolation filters already give a good approximation of the enhancement layer video, but have deficiencies in reconstructing the high frequency (HF) components. In [3], a sharpening filter is proposed which attempts to reconstruct the missing HF content in the base layer image. This refinement of HF content improves inter-layer prediction. However, the proposed filter does not utilize any learning based approach of HF reconstruction.

Image scale-up is a fundamentally ill-posed problem and interpolation can only approximate a blurry version of the actual high resolution (HR) image. Therefore, in the last decade, many approaches have been proposed which go beyond classical sampling theory. Recently the Sparse-Land Model [4] and its applications have shown good results in the field of single image super-resolution (SR). The general idea behind such SR methods relies on the fact that image data can be represented by a sparse linear combination of pretrained dictionary atoms. Yang et al. proposed to train dictionaries with coupled sparsity for low resolution and high resolution image patches in order to scale up a low resolution image [5], [6], [7]. Zeyde et al. used different image features which represent the HF content in images for the coupled training and reduced the processing time by applying a principal component analysis on the features [8]. All the SR methods show promising results in the case of single image SR but have not been used in the context of scalable video coding so far.

In this paper, a single image SR scale-up scheme for inter-layer prediction in scalable video coding is proposed. The method is based on learned basis functions and utilizes dictionary pairs with coupled sparsity for the SR task. The implementation was based on SPArse Modeling Software [9] and scalable HEVC reference software, SHM version 12.0. The remainder of the paper is organized as follows: Section II gives a short overview about the theory behind using dictionary learning (DL) and sparse coding for image representations. In section III, the dictionary learning and image scale-up SR-scheme used in this paper are detailed. Section IV documents the experimental setup and results. Finally, conclusions are drawn in section V.

II. Dictionary Learning for Image Representation

In this section, the basic principles of image representation via dictionaries and the notations are introduced. In order to represent an image $\mathbf{I}$ through a dictionary $\mathbf{D}$, the image is split into (usually overlapping) patches $\mathbf{X}$. In consequence, algorithms based on dictionary learning methods in the field of image processing are typically based on a patch-wise processing. A reconstruction of the full image is achieved by combining the patches and averaging over them in areas where patches overlap. Generally an image patch $\mathbf{x}$ can be approximated by a sparse linear combination of atoms which

IEEE VCIP'17, Dec. 10 - Dec. 13, 2017, St. Petersburg, USA
000-0-0000-0000-0/00/$31.00 ©2017 IEEE.

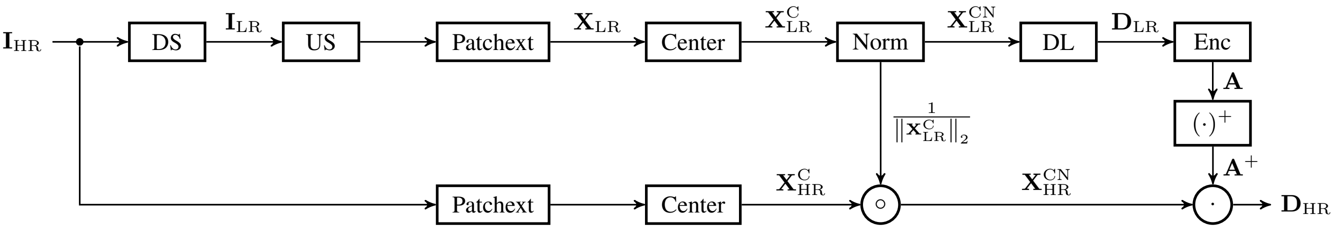
$\mathbf{I}_{\mathrm{HR}}$
DS
$\mathbf{I}_{\mathrm{LR}}$
US
Patchext
$\mathbf{X}_{\mathrm{LR}}$
Center
$\mathbf{X}_{\mathrm{LR}}^{\mathrm{C}}$
Norm
$\mathbf{X}_{\mathrm{LR}}^{\mathrm{CN}}$
DL
$\mathbf{D}_{\mathrm{LR}}$
Enc
$\mathbf{A}$
$(\cdot)^{+}$
$\mathbf{A}^{+}$
$\frac{1}{\|\mathbf{X}_{\mathrm{LR}}^{\mathrm{C}}\|_2}$
Patchext
Center
$\mathbf{X}_{\mathrm{HR}}^{\mathrm{C}}$
$\circ$
$\mathbf{X}_{\mathrm{HR}}^{\mathrm{CN}}$
$\cdot$
$\mathbf{D}_{\mathrm{HR}}$

Fig. 1. Coupled dictionary learning process. Note that the input is in general not only one but several training images. $\|\mathbf{X}_{\mathrm{LR}}^{\mathrm{C}}\|_2$ is a matrix containing patches with the l_2-norm of every path $\mathbf{X}_{\mathrm{HR}}^{\mathrm{C}}$ and $\circ$ denotes the Hadamard product. The output is given by the two dictionaries $\mathbf{D}_{\mathrm{LR}}$ and $\mathbf{D}_{\mathrm{HR}}$.

are the columns of a pretrained dictionary matrix $\mathbf{D}$, i.e.

$$\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha} \tag{1}$$

where only few entries in $\boldsymbol{\alpha}$ differ from zero. Note that the image patch $\mathbf{x}$ is the vectorized version of the actual patch. This sparse representation can be found solving

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1. \tag{2}$$

The parameter λ controls the sparsity of $\boldsymbol{\alpha}$ and has to be tuned to the needs of the application. Equation (2) can be solved e.g. by the LARS algorithm [4]. The dictionary $\mathbf{D}$ can be obtained by using a sufficient number of images as a training set, extracting n patches from the training set and solving the following equation,

$$\mathbf{D} = \arg\min_{\mathbf{D}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_i\|_1. \tag{3}$$

This could also be written as matrix factorization problem with $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ and $\mathbf{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_n]$

$$\mathbf{D} = \arg\min_{\mathbf{D}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_{\mathrm{F}}^2 + \lambda\|\mathbf{A}\|_{1,1}. \tag{4}$$

III. Single Image Scale-Up

A. Training of dictionaries with coupled sparsity for super-resolution

The general training procedure is depicted in Fig. 1. For the training of a coupled dictionary pair suitable for single image SR, it is necessary to generate corresponding training patch pairs $\mathbf{X}_{\mathrm{HR}}^{\mathrm{CN}}$ and $\mathbf{X}_{\mathrm{LR}}^{\mathrm{CN}}$ for the low and high resolution content. Therefore, the low resolution content is generated by a downsampling process denoted in the building block labeled with "DS" in Fig. 1. Note that this initial downsampling process is dependent on the application. For blind SR and visual reasonability this process might be unknown and can be modeled by some anti-aliasing filter and dropping of pixels. In the case of inter-layer prediction, it is useful to apply the same downsampling process as used for the generation of the base layer content. The filters used for the generation of the base layer content in our case are described in [10]. In order to obtain two dictionaries with the same size of atoms, the downsampled images $\mathbf{I}_{\mathrm{LR}}$ have to be upsampled again using some suitable interpolation method. For the results shown in this paper, the interpolation filters specified for interlayer prediction in SHVC being the same as used for subpixel motion compensation in HEVC [1] were used.

As pointed out in [8] it is sufficient to train the dictionaries for the high frequency (HF) components, since the low frequencies can be approximated by interpolation accurately. Note that we use centering (i.e. subtracting the mean) of the low resolution (LR) patches $\mathbf{X}_{\mathrm{LR}}$ as high-pass filter and not a Laplacian filter like in [8]. The reason for this lies in the fact that low resolution images generated according to [10] contain much more HF content compared to low resolution images obtained by e.g. Matlab's imresize function. Therefore, a Laplacian filter would strongly react on noise in the lower resolution images and introduce this noise to the dictionaries. Note that for centering, we assume that the LR patches and HR patches share the same mean value, as at the reconstruction side the mean values of the HR image are not available. Moreover, we apply a normalization step to the centered patches $\mathbf{X}_{\mathrm{LR}}^{\mathrm{C}}$ and $\mathbf{X}_{\mathrm{HR}}^{\mathrm{C}}$ in order to bring the patches to unit l_2-norm. As we want to reduce the chance of amplifying noise by this operation, the normalization is skipped if the l_2-norm of a patch is smaller than a threshold, i.e. in particular $\|\mathbf{x}_{\mathrm{LR}}^{\mathrm{C}}\|_2 < 0.1$. For the normalization, the HR patches $\mathbf{X}_{\mathrm{HR}}^{\mathrm{C}}$ are divided by the l_2-norms of $\mathbf{X}_{\mathrm{LR}}^{\mathrm{C}}$, as at the reconstruction side the original patches are not available and denormalization has to be done with low resolution patches.

Training a dictionary representing the HF of images at the lower resolution is straightforward solving (4) for image patches $\mathbf{X}_{\mathrm{LR}}^{\mathrm{CN}}$. This training is represented by the "DL" building block in Fig. 1 and as output the LR dictionary $\mathbf{D}_{\mathrm{LR}}$ is obtained. Consequently, the coefficients $\mathbf{A}$ can be calculated solving (2) for every training patch which is depicted as the "Enc" building block in Fig. 1. As coupled sparsity in both the LR and the HR dictionary is desired, the HR dictionary is calculated, such that the coefficient matrix $\mathbf{A}$ remains the same for both dictionaries. The following equation formulates that problem:

$$\mathbf{D}_{\mathrm{HR}} = \arg\min_{\mathbf{D}_{\mathrm{HR}}} \|\mathbf{X}_{\mathrm{HR}}^{\mathrm{CN}} - \mathbf{D}_{\mathrm{HR}}\mathbf{A}\|_2^2. \tag{5}$$

The closed form solution to that problem is given by pseudo inversion of the coefficient matrix $\mathbf{A}$:

$$\begin{aligned}\mathbf{D}_{\mathrm{HR}} &= \mathbf{X}_{\mathrm{HR}}^{\mathrm{CN}}\mathbf{A}^{+} &(6)\\ &= \mathbf{X}_{\mathrm{HR}}^{\mathrm{CN}}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1} &(7)\end{aligned}$$

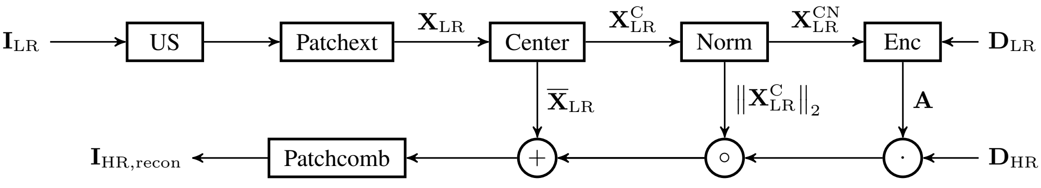
$\mathbf{I}_{\mathrm{LR}}$
US
Patchext
$\mathbf{X}_{\mathrm{LR}}$
Center
$\mathbf{X}^{\mathrm{C}}_{\mathrm{LR}}$
Norm
$\mathbf{X}^{\mathrm{CN}}_{\mathrm{LR}}$
Enc
$\mathbf{D}_{\mathrm{LR}}$
$\overline{\mathbf{X}}_{\mathrm{LR}}$
$\|\mathbf{X}^{\mathrm{C}}_{\mathrm{LR}}\|_2$
$\mathbf{A}$
$\mathbf{I}_{\mathrm{HR,recon}}$
Patchcomb
$+$
$\circ$
$\cdot$
$\mathbf{D}_{\mathrm{HR}}$

Fig. 2. Single image scale-up scheme. $\overline{\mathbf{X}}_{\mathrm{LR}}$ is a matrix containing patches with the mean intensity of every patch $\mathbf{X}_{\mathrm{LR}}$, $\|\mathbf{X}^{\mathrm{C}}_{\mathrm{LR}}\|_2$ is a matrix containing patches with the l_2-norm of every patch $\mathbf{X}^{\mathrm{C}}_{\mathrm{HR}}$ and $\circ$ denotes the Hadamard product.

This calculation of the dictionaries ensures that, once a sparse representation of an image in one of the dictionaries is found, the coefficients can be used to represent the image in the other dictionary, i.e. at a different resolution.

B. Scale-Up Scheme

Assuming that a pair of dictionaries with coupled sparsity has been trained as described in section III-A, single image SR can be performed according to the scheme depicted in Fig. 2. First the initial upsampling is performed in order to scale the input image to the size matching the trained dictionaries. For this purpose, the interpolation filters specified for inter-layer prediction in SHVC are used. Therefore, the dictionary learning based upsampling scheme could also be interpreted as a HF refinement of the upscaled base layer picture. In the sparse encoding stage, which is denoted as the building block labeled with "Enc", a sparse representation of the input in the LR dictionary is found. Since coupled sparsity for the dictionaries is assumed, this representation can directly be used to reconstruct the HF component of the HR image by multiplication of the coefficients $\mathbf{A}$ with the HR dictionary $\mathbf{D}_{\mathrm{HR}}$.

For the purpose of a reconstruction which covers the whole spectrum and not only the high frequencies, denormalization and decentering is applied. In the last step, the reconstruction HR image $\mathbf{I}_{\mathrm{HR,recon}}$ is generated by averaging over overlapping patch areas.

IV. Simulation Setup and results

A. Simulation Setup

In the following, the parameters used in the context of inter-layer prediction in scalable video coding are introduced. For the results reported in this paper, we considered the upscale factor between base layer and enhancement layer to be $f = 2$. The patch size was chosen to be $s_{\mathrm{p}} = 8 \times 8$ pixels and the patches overlap by 6 pixels in x and y direction. This overlap was chosen according to the idea of starting a new patch at a non-interpolated pixel position of the LR image. The dictionaries contain $K = 512$ atoms, which was observed to be sufficient for the application. The parameter λ introduced in section II was set differently for every rate point in order to make the upsampling method robust against the influence of coding artifacts. This is due to the fact that in the presence of ringing or blocking artifacts, those could be amplified by the scale-up scheme at a small λ. The reason for that lies in the fact that small values for λ promote less sparse but more accurate solutions of (2) and (4). In consequence, the artifacts could also be represented more accurately by the solution and even be amplified in the reconstruction phase. The Common Testing Conditions (CTC) for SHVC [11] define test cases with QP $\in \{22, 26, 30, 34\}$. Therefore, λ was set in dependency of the quantization parameter (QP), so that

$$\lambda = \begin{cases} 0.01, & \text{for QP} < 26 \\ 0.05, & \text{for } 26 \leq \text{QP} < 30 \\ 0.1, & \text{for } 30 \leq \text{QP} < 34 \\ 0.15, & \text{for } 34 \leq \text{QP} \leq 51. \end{cases} \tag{8}$$

Note that for every range of QPs a different pair of dictionaries has to be trained. The values in (8) were found to be close to optimum in a pre-analysis.

The training of the dictionaries was performed with small standard images which were also used in [8]. The resolution of the training images was in the range of roughly 80×80 to 430×350 pixels. The training set is different from the test set in terms of content and resolution. Note that it is not possible to perform a reasonable update (or even the training) of the dictionaries at the decoder side without spending rate for that, as the original images covering the high frequencies are not available at the decoder. Therefore, using pretrained dictionaries was preferred for this work.

In SHVC, interpolation filters for upsampling of the base layer picture are specified for the Luma and Chroma components. The proposed method replaces the interpolation filters in SHM 12.0 for the Luma component by the image scale-up process described in section III-B. The interpolation filters for the Chroma components were still used. Note that it is assumed that the dictionaries are available at the decoder side and are not coded into the bitstream. This assumption can be treated as valid since the dictionaries are not sequence dependent and in consequence it would be possible to store them permanently at the decoder.

The coding configurations and test sequences were taken from the CTCs for scalable video coding. The anchor was generated with SHM-12.0 reference software.

B. Results

In Tab. I BD-rate savings, calculated by cubic spline interpolation between the rate points and integration over the difference, are depicted. The measured values correspond to Y-PSNR over rate plots. The results show clearly that rate savings can be achieved with the DL scale-up scheme for

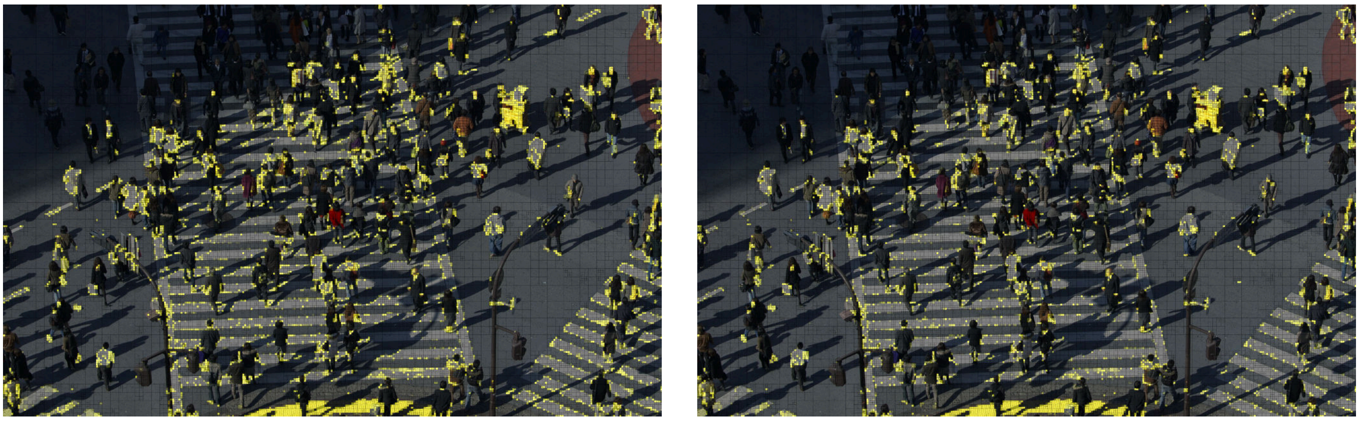

Fig. 3. First frame of PeopleOnStreet. CUs consuming more than 1 bit per pixel at QP = 22 on average are marked in yellow. Left: reference SHM 12.0, right: SHM 12.0 with DL SR inter-layer prediction

	AI	RA	LD
BQTerrace	−2.25 %	−0.89 %	−0.33 %
BasketballDrive	−1.28 %	−1.37 %	−1.19 %
Cactus	−1.88 %	−1.04 %	−0.59 %
Kimono	−0.58 %	−0.59 %	−0.35 %
ParkScene	−1.26 %	−1.05 %	−0.52 %
PeopleOnStreet	−6.15 %	−5.93 %	−5.13 %
Traffic	−3.02 %	−2.05 %	−1.03 %
AVG	−2.35 %	−1.85 %	−1.31 %

TABLE I
BD-RATE SAVINGS FOR DIFFERENT CODING CONFIGURATIONS AT AN UPSCALE FACTOR OF 2. THE ANCHOR WAS SHM 12.0

inter-layer prediction. Averaged rate savings of 2.35 % for All Intra (AI), 1.85 % for Random Access (RA) and 1.31 % for Low Delay (LD) compared to the standardized scalable video coding extension of HEVC are observed. These rate savings are related to the fact that the proposed upscaling method performs better at high frequencies which leads to lower coding costs for the residual of the enhancement layer. Note that the coding gain is highly dependent on the coded video sequence due to different characteristics. For example, the method performs worst for the sequence Kimono, which is known to contain a low amount of HF content. The best performance is observable for PeopleOnStreet. This is related to the high amount of sharp edges and structures in the video sequence. In Fig. 3, the first frame of PeopleOnStreet is shown with CUs consuming more than 1 bit per pixel on average marked in yellow. It can be seen that in areas of edges the amount of yellow marked CUs is less for the proposed inter-layer prediction method than for the reference. Moreover, the results show that when temporal prediction is enabled, the impact of the improved inter-layer prediction becomes less important, as inter prediction from the enhancement layer frames is more efficient than inter-layer prediction in many cases. Last, it should be mentioned, that the complexity of DL based inter-layer prediction clearly exceeds the complexity of filtering.

V. Conclusion

In this paper, an approach for inter-layer prediction based on dictionary learning SR is presented. It is shown that state of the art interpolation filters are clearly outperformed by the proposed method. However, the method presented in this paper has the potential for further improvements. For example, the inclusion of other temporal reference pictures from the base layer in the SR algorithm could give a better approximation of the enhancement layer pictures. Thus, combining dictionary learning and motion estimation based SR approaches could be of interest. In conclusion, the investigations and results of this paper show that room for improvement exists in inter-layer prediction of SHVC.

References

[1] B. Bross, W. Han, J. Ohm, G. Sullivan, Y. Wang, and T. Wiegand, *High Efficiency Video Coding (HEVC) text specification draft 10 (for FDIS and Last Call)*, JCTVC-L1003, ISO/IEC JTC1/SC29/WG11 Std. JCTVC-L1003, Jan. 2013.

[2] J. Chen, K. Rapaka, X. Li, V. Seregin, L. Guo, M. Karczewicz, G. V. D. Auwera, J. Sole, X. Wang, C. Tu, Y. Chen, and R. Joshi, "Scalable video coding extension for hevc," in *Data Compression Conference (DCC), 2013*, March 2013, pp. 191–200.

[3] M. Sychev, S. Ikonin, and V. Anisimovskiy, "Sharpening filter for interlayer prediction," in *Visual Communications and Image Processing Conference, 2014 IEEE*, Dec 2014, pp. 470–473.

[4] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[5] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8.

[6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov 2010.

[7] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3467–3478, Aug 2012.

[8] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[10] J. Chen, J. Boyce, Y. Ye, M. Hannuksela, and G. Barroux, "SHVC test model 11 (SHM 11) introduction and encoder description," JCTVC, 22nd Meeting, Geneva, Tech. Rep. JCTVC-V1007, oct 2015.

[11] Y. Chen, G. Tech, K. Wegner, and S. Yea, "Common SHM test conditions and software reference configurations," JCTVC, 17th Meeting, Valencia, Tech. Rep. JCTVC-Q1009, Mar. 2014.