Adaptive Down-Sampling and Super-Resolution for Additional Video Compression

Bin Zhao[1] and Jianmin Jiang[2](✉)

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China
woxintaoxiang@outlook.com

[2] Research Institute of Future Media Computing, Shenzhen University, Shenzhen, China
jianmin.jiang@szu.edu.cn

Abstract. While almost all down-sampling based video codecs gain additional compression at the expense of image degradation, we set a good example of achieving both large compression and even better reconstruction quality. Such progress is realized by: (i) minimizing the introduction of information loss with a proposed decomposition-based adaptive down-sampling method so that more reserved pixels can be allocated to image details where human visual perception is more sensitive. Specifically, a modified content complexity measurement is put forward and the optimum down-sampling rate is adaptively selected with a customized formula; (ii) maximizing the information compensation via a content-adaptive super-resolution algorithm, which is accelerated and optimized by two stages of pruning to select the closest correlated dictionary pairs. Extensive experiments support that, by using prevailing H.264 codec as benchmark, the proposed scheme achieves 5 times more of additional compression and the reconstruction quality outperforms other state-of-the-art approaches, and even better than decoded non-shrunken frames in human visual perception.

Keywords: Video compression · Decomposition · Super resolution · Sparse representation

1 Introduction

Currently, the ubiquitous application of digital mobile video and high-definition (HD) visual enjoyment suffer severe bottleneck due to limited bandwidth. Though off-the-shelf codecs, such as the prevailing H.264/MPEG-4 AVC [1] and one of its potential successors, the H.265/HEVC (High Efficiency Video Coding) [2] have provided sharp compression, further shrinkage is still required. Accordingly, the scaling based coding schemes, which down- and up-sample video frames respectively prior and posterior to generic codecs, stand out as a feasible division of ongoing research for compression improvement [3–6]. However, for pertinent literature, the up-sampling or essentially the super resolution methods have been actively researched to compensate the discarded high frequency details while little efforts were made to optimize the information loss that the down-sampling process introduced. Most approaches simply applied a fixed down-sampling rate to the whole frames and neglected the heterogeneity of human visual system

L.A. Maglaras et al. (Eds.): INISCOM 2016, LNICST 188, pp. 111–121, 2017.
DOI: 10.1007/978-3-319-52569-3_10

in the tolerance of content information loss. As [7] suggested, to complex regions with more details, human visual perception is more susceptible, thus less pixel loss should be brought in. For this purpose, we proposed a decomposition-based adaptive down-sampling scheme, which initially executes two successive decomposition, temporal and spatial, to divide the input video into a number of fragment sequences. Each of them contains a certain level of complexity and is appropriate to adopt diverse down-sampling rate. Specifically, to achieve the best possible trade-off between compression effectiveness (reconstruction quality at the decoder end) and compression efficiency (additional compression ratios), we developed a local-homogeneity-based global metric (LHGM) as decomposition criterion and paved the way for the selection of optimum down-sampling rate. The resulting decomposed and down-sampled sub-sequences can be encoded separately by any existing codecs. And the decoded counterparts will then be up-sampled and pasted together with the along coded side information.

At the decoder, resolution enhancement and detail compensation critically count on super-resolution (SR) reconstruction techniques, among which the dictionary- or example-based learning methods protrude as the most active area over recent years. By using learned co-occurrence prior knowledge between low resolution (LR) and high resolution (HR) image patches, the learning-based methods effectively overcome a variety of deficits in other SR techniques, for instance, the over-smoothness or ringing and jagged artifacts of interpolation-based methods [8] and the requirement of multiple aliasing frames of the same scene in multi-frame methods [9]. A typical dictionary-based example is the Neighbor Embedding (NE) method, which utilizes the local geometry similarity between LR image patches and their HR counterparts [10]. Recent Nonnegative Neighbor Embedding (NNE) [11] and Anchored Neighborhood Regression (ANR) [12] both stemmed from this hypothesis. Sparse Coding (SC) represents another major direction of dictionary-based learning methods. Yang et al. [13] exploited joint sparse representation for the input LR and expecting HR image patches based on a pair of over-complete dictionaries and produced better SR images with fine details. Zeyde et al. [14] built upon his framework and improved the execution speed by modifying the training approaches. However, from extensive investigation, we found that the majority of dictionary-based algorithms primarily focused on the selection of the nearest neighbors in NE or the representation of sparse signals in SC, while the preparation for the dictionaries was often ignored, which make space for our exploration in this aspect.

Besides, due to the fact that receivers at the decoder end tend to evaluate video quality based on direct visual impact rather than computed signal-to-noise ratio that compared with the pre-encoding references, it' s more reasonable to assess reconstructed image quality according to human visual system [7]. Given that, we adopted the blind/no-reference image quality assessment model (BRISQUE) [23] in reconstruction evaluation, which was demonstrated statistically better than the full-reference peak signal-to-noise-ratio (PSNR) and structural similarity index (SSIM) in both computational efficiency and the correlation with human visual characteristics.

The rest of the paper proceeds as follows: Sect. 2 is a fine description of the proposed quadtree-based decomposition and adaptive down-sampling methods. Section 3 theoratically formulized the content-adaptive sparse representation of super-resolution. Section 4 reports our experimental results and Sect. 5 provides concluding remarks and some ideas for future work.

2 Decomposition-Based Down-Sampling

The design of our decomposition-based down-sampling scheme was in view of the non-uniform distribution of video content and the benefit of applying flexible down-scaling rates to complexity-varied regions. Meanwhile, considering the computational cost, we assumed that content distribution across video frames like color, texture and motion remains relatively stable within one scene and appropriate to apply a unified decomposition pattern in accordance with the first frame or Intra-frame.

To achieve above ideas, we first utilized a content change detection similar to shot cut technique [15] to temporally divide the input video into several scenes, or so called V-units [16] (shown as Fig. 1 Temporal Decomposition). In each V-unit, the frame number was limited (not more than twenty in our case) to make sure that a certain level of content-consistency is maintained. Then we implemented a quad-tree decomposition on the first frame of each V-unit. In this process, a content complexity measurement is required for determining whether a frame patch should be divided.

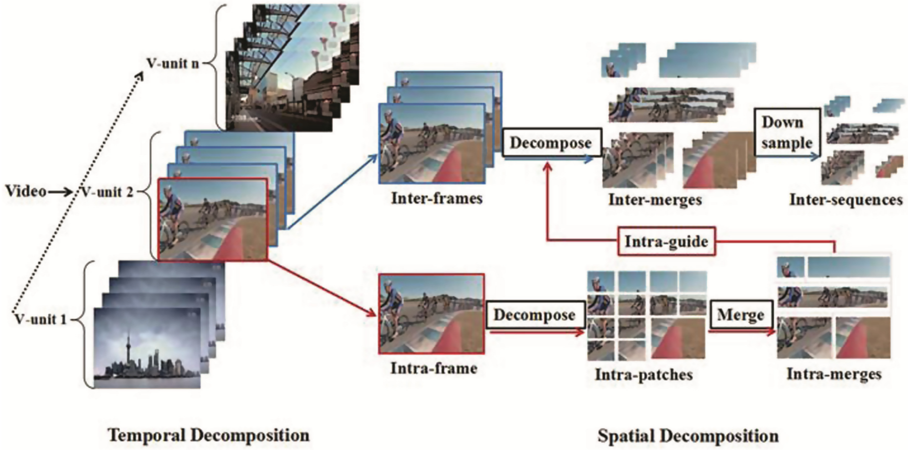

Fig. 1. Illustration of decomposition-based adaptive down-sampling

Existing complexity measurements [17–20], such as the variance-based [17] and edge-based approaches [20], were primarily designed for the selection of partition mode, which made them only suitable for the calculation of coding blocks. In that, directly applying these formula to large frame patches would lead to astronomical figures. On the contrary, MSE (mean square error) calculation was free of scale limitation while its result value lack reliability on discrepant content characteristics and lack stability on the change of luminance or chrominance even when the image textures remain consistent. To solve the dilemma, we proposed a local-homogeneity-based global metric (LHGM) in formula (1) to accurately reflect image details. Importantly, our algorithm is unconstrained by image scales. At the same time, the result is unaffected by the variation of image size.

$$LHGM(p) = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[g(i,j) - g^{s}(i,j)\right]^{2} \tag{1}$$

Here *p* represents a frame patch with the size of $m \times n$, *g(i,j)* is the gray-level at pixel *(i,j)*, and $g^s(i,j)$ is the averaged gray-level in an $s \times s$ window centered at *(i,j)*. The size of *s* depends on the requirement of detail levels (in our application, it is empirically set at 5 pixel).

To illustrate our quadtree-based hierarchical decomposition, we use *p (Lp, Ip, Dp)* to represent a patch node in the tree structure. Here *Lp* indicates the decomposition level of *p*, *Ip* indexes *p* at that level, and *Dp* gives the down-sampling rate that to be applied at the current patch. By this definition, the original version of each first frame, like the Intra-frame in V-unit 2 (Fig. 1 Spatial Decomposition) is labeled as a root node ($Lp = 0$). Next, we decompose it into four equal parts via a quad-tree decomposition and calculate each of their content complexity with formula (1). If any resulting LHGM value exceeds a predefined threshold (70 db), indicating the existence of detailed fragment(s), it will be replaced by the four partitions. Otherwise, the Intra-frame will be treated as a leaf node and processed as a whole. Similarly, each partition with an over-threshold LHGM value will be further decomposed iteratively until either its four sub-patches are all smaller than the threshold or the highest level K (limited to video resolution and minimum coding size) is reached. For implementation efficiency, adjacent leaf patches of the final partitions (the Intra-patches in Fig. 1.) that belong to the same complexity range (with the same *Dp* value) are merged together so that down-sampling can be operated integratedly later. The Intra-merges in Fig. 1 displays the final decomposition pattern of V-unit 2 and succeeding Inter-frames can just follow this pattern. Such process is named as Intra-Guide.

In general, patches at deeper tree levels contains more details and ought to be compressed with larger down-sampling rates. Exception exist when a smooth area is separated out due to the high LHGM value of its brother nodes. So during the selection of adaptive down-sampling rate *Dp,* we synthetically considered both the quad-tree level *Lp* and the the patch's content complexity. Exact formula is given below:

$$D_p = \begin{cases} \partial^{K-L_p+1} & LHGM(p) < threshold \\ \partial^{K-L_p} & LHGM(p) > threshold \end{cases} \tag{2}$$

Where ∂ is a base number, deciding the scaling extent and can be adjusted according to the limitation of bandwidth or the requirement of reconstruction quality.

3 Content-Adaptive Sparse Representation of Super-Resolution

Based on sparse signal representation, research on image statistics suggested that image patches can be well-represented as a sparse linear combination of elements from an appropriately chosen over-complete dictionary [13] as expressed below:

$$I = D\alpha \tag{3}$$

Where $I \in R^n$ is a given image, $\alpha \in R^k$ is a sparse vector with very few ($<< k$) nonzero elements, and $D \in R^{n\times k}$ is an over-complete dictionary of k prototype signal-atoms. Later, theoretical results from compressed sensing [21] demonstrated that under mild conditions, the sparse representation of a HR patch can be correctly recovered from its down-sampled version. Inspired by above facts, Yang et al. [13] designed their SR approach in following steps: (i) given a set of HR images, jointly train two dictionaries D_H and D_L for original patches and their down-sampled version; (ii) represent a LR input patch I_L by:

$$\min\|\alpha\|_1 \quad s.t.\left\|F(D_L\alpha) - F(I_L)\right\|_2^2 \le \varepsilon \tag{4}$$

Where F is a feature extraction operator, which provides a constraint on how closely the α approximates D_L; (iii) apply the LR sparse representation α with D_H to generate a super resolved image patch.:

$$I_H = D_H\alpha \tag{5}$$

Since the reconstruction of I_H is based on those HR training images, whose content has the highest level of similarity to I_H, it is crucial that the two training dictionaries D_H and D_L be optimized. To this end, we proposed to analyze the input LR image I_L and extract its content features to select corresponding HR training images in the dictionary preparation. In this way, the super-resolution process can be made adaptive to the content of input video frames, and hence achieve the advantage that their reconstruction quality can be further improved. The problem of training the two dictionaries D_H and D_L can be formulated as:

$$D_H = \underset{\{D_H,C\}}{\arg\min} \left\|f_H\left(T, I_L\right) - D_H C\right\|^2 + \lambda\|C\|_1 \tag{6}$$

$$D_L = \underset{\{D_L,C\}}{\arg\min} \left\|f_L\left(T, I_L\right) - D_L C\right\|^2 + \lambda\|C\|_1 \tag{7}$$

Where $C \in R^{r\times m}$ is a coefficient matrix, r is determined by the dimension of selected content features, the l_1 norm $\| C \|_1$ is used to enforce the extent of sparse, and T is the set of training image candidates. Finally, $f_H(T, I_L)$ and $f_L(T, I_L)$ are the sets of optimized HR and LR image patches, the elements of which are derived through two stages of pruning process.

For the first stage of pruning, we applied a texture descriptor [22], which is based on three edge patterns, to facilitate the similarity based selection. The basic principle is to examine two DCT coefficients $X(0, 1)$ and $X(1, 0)$ for every block of 8×8 pixels with any DCT-based video compression technology and produce the following block-edge patterns (BEP), Where λ is a threshold to control the number of edge patterns:

$$BEP = \begin{cases} no\_edge & if \max(|X(1,0)|, |X(0,1)|) < \lambda \\ vertical\_edge & if\ |X(0,1)| \geq |X(1,0)|) \\ horizontal\_edge & otherwise \end{cases} \tag{8}$$

The texture descriptor is then constructed via a histogram-based approach to pre-select those training images in compressed domain, that are likely to produce good match for training dictionaries [22].

The second stage of pruning is carried out at the image patch level, where direct match among detailed content features out of SIFT is implemented. Specific process is described as follows:

$$e_H(T, I_L) = P_H\left(\underset{x_i \in T}{\arg\min} \{d[SIFT(x_i), SIFT(I_H)]\}\right) \tag{9}$$

$$e_L(T, I_L) = P_L\left(\underset{x_i \in T}{\arg\min} \{d[SIFT(x_i), SIFT(I_L)]\}\right) \tag{10}$$

Where $P_H(.)$ and $P_L(.)$ are two operators that segment the image into HR patches and LR patches, respectively. $d(.)$ is the Euclidean distance, which is used to measure the similarity between the SIFT descriptors of the training image candidate x_i and the input image patches I_L or I_H.

4 Experimental Results

To evaluate the proposed additional compression algorithm, we carried out extensive experiments on video sequences with various sizes, mostly VGA (640 × 480p), SD (1280 × 720p) and HD (1920 × 1080p). The down-sampling rates that we adopted ranged from 2 to 8, depending on the level of content complexity. For each video frame, both of their original and down-sampled versions were encoded by JM codecs with 5 QP parameters (12, 18, 24, 30, 36). After the decoder, those down-scaled frame patches were enlarged with proposed SR method and 4 additional state-of-the-art SR methods that we mentioned in Introduction. Their eventual recovery quality were compared together with h.264 directly reconstructed, non-downsampled frames in terms of BRISQUE value.

Table 1. summarized the experiment results from 3 video samples with the size of VGA, SD and HD separately. And each BRISQUE value was an average of 120 video frames. As shown in Table 1, all of the super resolution methods provide competitive performance, among which our proposed is obviously the best, even outperforms the h.264 directly decoded non-shrinked ones. Especially in case of large QP values where transformation coefficients are quantified with coarse parameters, our super resolution method exhibits remarkable compensation effect for high frequency information loss. To more clearly demonstrate the promotion of restoration quality with our method, we appended the multiple spline curves of QP vs. BRISQUE value for the listed methods in Fig. 2.

Table 1. BRISQUE value of control methods and our proposed.

BRISQUE	QP	h.264	Proposed	Zeyde	ANR	NE	NNLS
VGA	12	0.8697	0.8379	0.7690	0.7647	0.7641	0.7669
	18	0.8380	0.8389	0.7528	0.7519	0.7500	0.7531
	24	0.7965	0.8241	0.7301	0.7269	0.7258	0.7294
	30	0.7539	0.8064	0.7073	0.7039	0.7023	0.7060
	36	0.6460	0.7301	0.6176	0.6161	0.6162	0.6185
SD	12	0.5773	0.6130	0.5294	0.526	0.5341	0.5282
	18	0.6037	0.6128	0.5358	0.5335	0.535	0.5345
	24	0.5824	0.6095	0.5296	0.5289	0.5298	0.5296
	30	0.5295	0.5887	0.5064	0.5053	0.5070	0.5067
	36	0.4671	0.5413	0.4507	0.4512	0.4504	0.4514
HD	12	0.6375	0.6877	0.6053	0.6007	0.6042	0.6138
	18	0.6425	0.6859	0.6072	0.6035	0.6087	0.6180
	24	0.6101	0.6805	0.5993	0.5947	0.6005	0.6086
	30	0.5560	0.6526	0.5699	0.5648	0.5705	0.5779
	36	0.4839	0.5897	0.5041	0.4995	0.5034	0.5096

Besides above quantitive valuation, figurate examples (the original frame, h.264 decoded, proposed reconstructed and one best example of other four contrast methods reconstructed frame) are presented in Fig. 3. It can be seen that our proposed is obviously the clearest.

The original size of 3 video sequences and the size of their bitstream that respectively compressed by h.264 codec and our scheme are listed in Table 2. Compared with h.264, our proposed gain an add-on compression of over 5 times in average, indicating that if our proposed codec is applied to telecommunication, the transfer speed will be improved by more than five times. In terms of storage, application of the proposed codec will give us additional savings of about five sixths of the overall storage space.

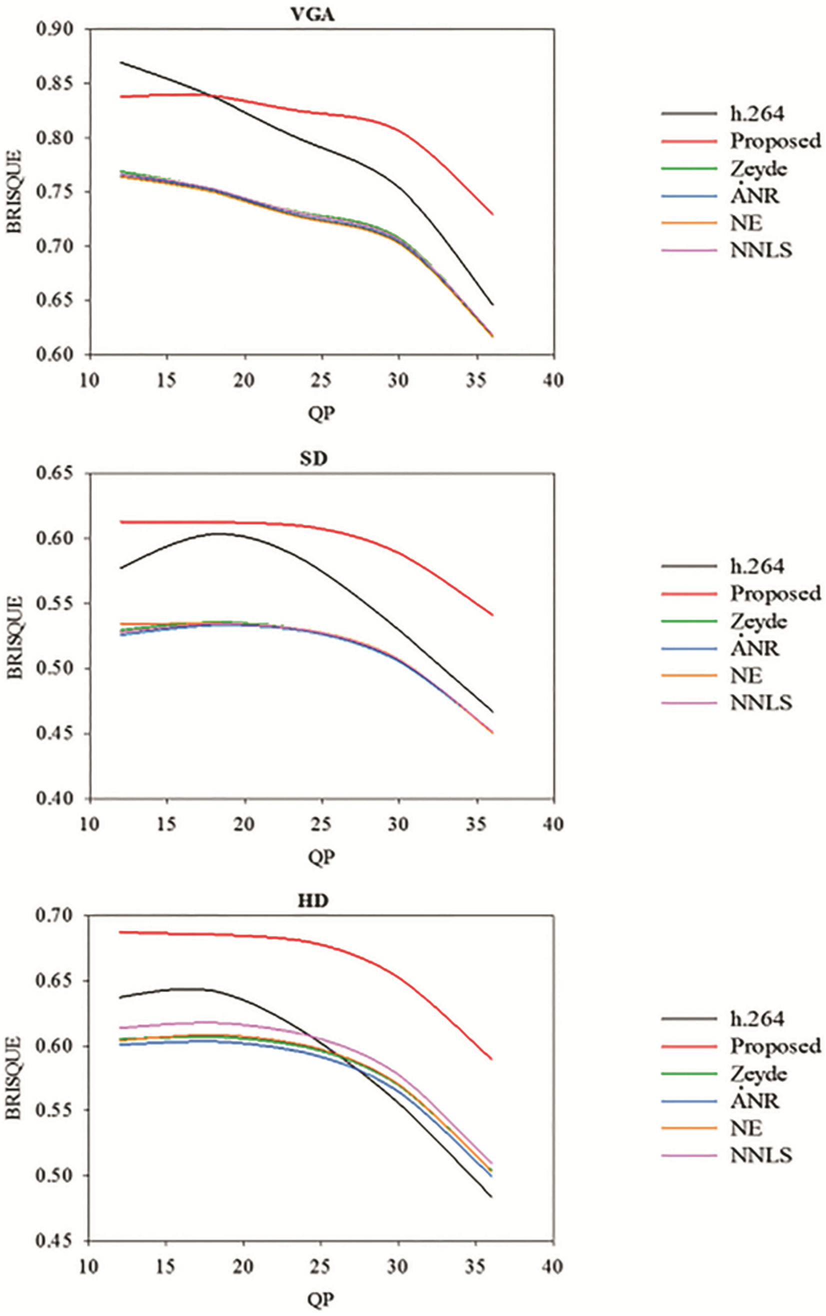

Fig. 2. BRISQUE vs. QP value of h.264, our proposed and 4 control methods

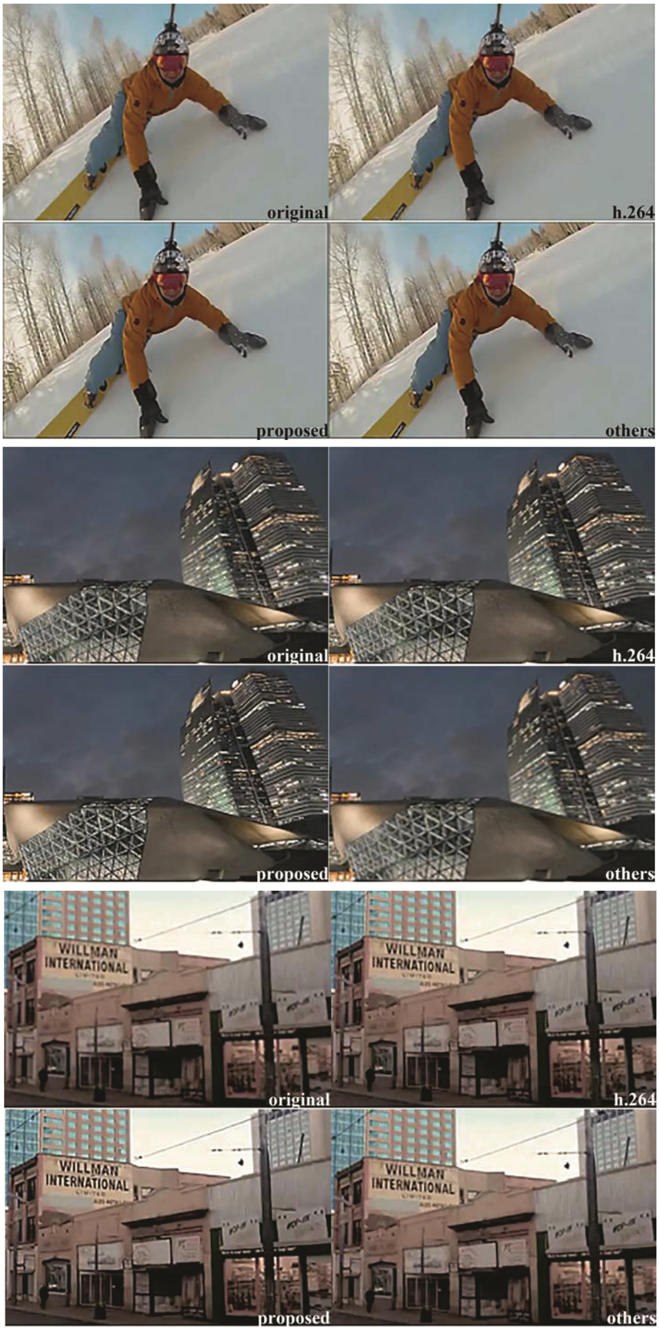

Fig. 3. Illustration of reconstruction frames

Table 2. Averaged data size of original video, compressed bitstream via h.264 and our proposed scheme, as well as the add-on compression ratio.

Data size	Original (KB)	QP	h.264 (KB)	Proposed (KB)	Add-on
VGA	20172	12	700	217	3.23
		18	389	120	3.24
		24	190	61	3.14
		30	86	28	3.01
		36	37	12	3.00
SD	37800	12	1132	126	8.98
		18	623	72	8.65
		24	315	37	8.51
		30	148	18	8.22
		36	64	8	8.00
HD	60723	12	2632	542	4.86
		18	1229	284	4.33
		24	548	139	3.95
		30	220	70	3.15
		36	107	36	2.99

5 Conclusion

In this paper, we designed an adaptive video compression technique on top of existing codecs for achieving additional contraction, accelerated transformation and also excellent reconstruction quality. The performance discrepancies in aspects of additional compression ratio and reconstruction quality may be associated with differential video resolutions, content complexity or other video characteristics, which remain to be researched in future work.

References

1. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the h.264/avc standard. IEEE Trans Circ. Syst. Video Technol. **17**(9), 1103–1120 (2007)
2. Sullivan, G.J., Ohm, J., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circ. Syst. Video Technol. **22**(12), 1649–1668 (2012)
3. Shen, M., Xue, P., Wang, C.: Down-sampling based video coding using super-resolution technique. IEEE Trans Circ. Syst. Video Technol. **21**(6), 755–765 (2011)
4. Chang, P.C.: Adaptive down-sampling video coding. In: Proceedings of SPIE - The International Society for Optical Engineering, vol. 24(11), pp. 1957–1968 (2010)
5. Nguyen, V.A., Tan, Y.P. Lin, W.: Adaptive downsampling/upsampling for better video compression at low bit rate. In: IEEE International Symposium on Circuits and Systems, pp. 1624–1627 (2008)
6. Alfred, M.B., Elad, M., Kimmel, R.: Down-scaling for better transform compression. IEEE Trans. Image Process. **12**(9), 1132–1144 (2003)

7. Jiang, J.: A low-cost content-adaptive and rate-controllable near-lossless image codec in dpcm domain. IEEE Trans. Image Process. **9**(4), 543–554 (2000)
8. Li, X., Orchard, M.T.: New edge directed interpolation. IEEE Trans. Image Process. **10**(10), 1521–1527 (2001). A Publication of the IEEE Signal Processing Society
9. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Trans. Image Process. **13**(10), 1327–1344 (2004)
10. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: IEEE Computer Society Conference on Computer Vision & Pattern Recognition, pp. 275–282. IEEE Computer Society (2004)
11. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC (2012)
12. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: IEEE International Conference on Computer Vision (ICCV), pp. 1920–1927 (2013)
13. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010). A Publication of the IEEE Signal Processing Society
14. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Boissonnat, J.-D., Chenin, P., Cohen, A., Gout, C., Lyche, T., Mazure, M.-L., Schumaker, L. (eds.) Curves and Surfaces 2010. LNCS, vol. 6920, pp. 711–730. Springer, Heidelberg (2012). doi:10.1007/978-3-642-27413-8_47
15. Ren, J., Jiang, J., Chen, J.: Shot boundary detection in MPEG videos using local and global indicators. IEEE Trans. Circ. Syst. Video Technol. **19**(19), 1234–1238 (2009)
16. Ren, J., Jiang, J.: Hierarchical modelling and adaptive clustering for real-time summarization of rush videos. IEEE Trans. Multimed. **11**(5), 906–917 (2009)
17. Min, B., Cheung, R.C.C.: A fast cu size decision algorithm for the HEVC intra encoder. IEEE Trans. Circ. Syst. Video Technol. **25**(5), 892–896 (2015)
18. Zhao, T., Wang, Z., Kwong, S.: Flexible mode selection and complexity allocation in high efficiency video coding. IEEE J. Sel. Topics Signal Process. **7**(6), 1135–1144 (2013)
19. Cho, S., Kim, M.: Fast cu splitting and pruning for suboptimal cu partitioning in HEVC intra coding. IEEE Trans. Circ. Syst. Video Technol. **23**(9), 1555–1564 (2013)
20. Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y.: Soft edge smoothness prior for alpha channel super resolution. In: IEEE Conference on Computer Vision & Pattern Recognition, pp. 1–8 (2007)
21. Kondo, S.: Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theor. **54**(5), 2210–2219 (2008)
22. Jiang, J., Armstrong, A., Feng, G.C.: Direct content access and extraction from JPEG compressed images. Pattern Recogn. **35**(11), 2511–2519 (2002)
23. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: a natural scene statistics approach in the DCT domain. IEEE Trans. Image Process. **21**(8), 3339–3352 (2012)