Neurocomputing 162 (2015) 180–190

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Model-based low bit-rate video coding for resource-deficient wireless visual communication

Xianming Liu [a,*], Xinwei Gao [a], Debin Zhao [a], Jiantao Zhou [b], Guangtao Zhai [c], Wen Gao [d]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[b] Faculty of Science and Technology, University of Macau, E11 Avenida da Universidade, Taipa, Macau, China
[c] Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China
[d] School of Electronic Engineering and Computer Science, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 9 October 2014
Received in revised form
16 January 2015
Accepted 24 March 2015
Communicated by Y. Yuan
Available online 14 April 2015

Keywords:
Wireless visual communication
Low bit-rate video coding
Low complexity

ABSTRACT

In this paper, an effective low bit-rate video coding scheme is developed to realize state-of-the-art video coding efficiency with lower encoder complexity, while supporting standard compliance and error resilience. Such an architecture is particularly attractive for application scenarios involving resource-deficient wireless video communications. At the encoder, in order to increase resilience to channel error, multiple descriptions of a video sequence are generated in the spatio-temporal domain by temporal multiplexing and spatial adaptive downsampling. The resulting side descriptions are interleaved with each other in temporal domain, while still with conventional square sample grids in spatial domain. As such, each side description can be compressed without any change to existing video coding standards. At the decoder, each side description is first decompressed, and then reconstructed to the original resolution with the help of the other side description. In this procedure, the decoder recovers the original video sequence in a constrained least squares regression process, in which 2D or 3D piecewise autoregressive model is adaptively chosen according to different predictive modes. In this way, the spatial and temporal correlation is sufficiently explored to achieve superior quality. Experimental results demonstrate that the proposed video coding scheme outperforms H.264/AVC and other state-of-the-art methods in rate–distortion performance at low bit-rates and achieves superior visual quality at medium bit rates as well, while with lower encoding computational complexity.

1. Introduction

In recent years, low-cost devices such as CMOS cameras that are able to ubiquitously capture video content from the environment have appeared in almost all small wireless mobile devices, such as smart-phones and tablet PCs. Furthermore, recent developments in sensor networking have encouraged the use of video sensors in these networks, which has fostered the development of Wireless Video Sensor Network (WVSN) [1–5]. Following the trends in low-power processing, wireless networking, and distributed sensing, WVSN has developed as a new technology with a number of potential applications, ranging from mobile multimedia, security monitoring, and advanced health care delivery to emergency response.

WVSN will be composed of inter-connected, battery-powered miniature video sensors. WVSN is usually mission driven and application specific, therefore it must operate under a set of unique constraints and requirements. The main concern for WVSN is that the energy provisioned for a video sensor is not expected to be renewed throughout its mission because sensor nodes may be deployed in a hostile or unpractical environment. And the bandwidth of wireless channels is usually limited. At the same time, it is necessary to provide some error-resilient mechanism against instability of wireless channels. More specially, there are several factors that mainly influence the design of a WVSN:

- *Compression efficiency*, *encoder complexity and bandwidth*: An effective video compression scheme can significantly reduce the amount of video data to be transmitted, which in turn saves a significant amount of energy in data transmission. However, more effective video compression methods often require higher computational complexity. These two conflicting effects imply that in practical system design there is always a tradeoff among compression efficiency, encoding complexity and bandwidth.
- *Resiliency to channel errors*: Wireless channels are unstable and noisy. Therefore, the source coder should provide some mechanism for robust and error-resilient coding of source data.

There exists a vast literature on video coding techniques. In traditional hybrid video compression standards, such as MPEG-2

* Correspondingauthor.
E-mail address: xmliu.hit@gmail.com (X. Liu).

http://dx.doi.org/10.1016/j.neucom.2015.03.054
0925-2312/

[6], H.264/AVC [7] and HEVC, consecutive frames are encoded jointly to achieve maximum coding efficiency, which are based on the idea of predictive coding to exploit spatio-temporal correlations. Although achieving the state-of-the-art rate-distortion performance, they may not be suitable for low-cost video sensors since predictive coding requires complex encoders and entails high energy consumption. Besides, the predictive coding system causes inter-frame dependency in decompression, which results in error propagation for an error-prone channel. Another approach in conventional wisdom is distributed video coding (DVC) [8–10], which shifts the complexity to the decoder end to allow the use of simple encoders. At the same time, DVC has an inbuilt robustness to channel losses because there is a duality between distributed source coding and channel coding. Clearly, DVC is very promising for WVSN. However, despite years of intensive research on DVC, the current systems still fail to meet the compression efficiency of their predictive coding counterparts [11–13]. Moreover, the feedback channel is required in existing mainstream DVC schemes, which will introduce delay and therefore is not suitable for practical WVSN applications.

In this paper, we investigate the sparse sampling based approach for wireless video communication. The fact that natural videos have an exponentially decaying power spectrum in spatial domain and strong correlations in temporal domain suggests the possibility of interpolation-based compact representation of video signals. We find that sparse sampling naturally fulfills the role of encoder because it greatly reduces the amount of data. To achieve maximum coding efficiency, the downsampled data should be further compressed. For this purpose we choose a uniform downsampling scheme to generate conventional square sample grids with smaller size. In this way, the information needed to be compressed and transmitted is significantly reduced, so that the proposed scheme can greatly reduce the encoder complexity, and naturally prolong the operational lifetime of video sensors. Similar to DVC, the proposed scheme provides an asymmetric video codec design, which shifts the associated computation burdens to the decoder. In this system design, the heavy-duty video decoding can be performed by powerful computers or perceivably in near future by cloud computing.

On the other hand, the needs for wireless-network-aware coding techniques to mitigate the problem of packet losses have generated much interest in multiple description coding. In the more simple architecture, MDC of a video source consists in generating two equivalent importance data streams that, all together, carry the input information. At the receiver side, when both the descriptions are available a high quality video is reconstructed. If only one bit-stream is available at the decoder end, a poorer but acceptable quality reconstruction is obtained. MDC has emerged as a promising approach to enhance the error resilience of a video delivery system. However, most of the existing video MDC techniques [14-17,19,27-29] are not compliant to existing video coding standards, either being completely different approaches or requiring a significant degree of modifications to an existing standard. Moreover, since introducing redundancy for error resilience, the compression efficiency of MDC based schemes is usually lower than the conventional video coding standard. A natural question is whether combing compact signal representation with MDC can be made a practical and competitive solution for wireless video communications. In this paper we will give an affirmative answer to the above question.

The rest of this paper is organized as follows. In Section 2, we overview the related work. Section 3 presents the framework of the proposed scheme. Sections 4 and 5 detail the main contribution of this paper: mode-dependent soft-decoding via constrained least squares regression. Section 6 presents the experimental results and comparative studies. Section 7 concludes the paper.

2. Related work

In the literature, many interpolation-based low bit-rate image/video compression algorithms have been proposed. In [29], an interpolation-dependent downsampling method was proposed to hinge the interpolation to the downsampling process. In [18], Shen et al. proposed a downsampling based video coding scheme, where an example based super-resolution technique is employed to restore the downsampled frames to their original resolutions. This work needs an offline training set with different video resolution. In some practical cases, the training set is not available. In [20], an adaptive downsampling mode decision in the encoder was proposed. The modes including different directions and sizes can be determined by the features of block contents. Ref. [30] proposes to find the optimal downsampling ratio that balances the distortions caused by downsampling and coding, thus achieving the overall optimal RD performance. However, this method will introduce heavy computation complexity at the encoder side to perform complex rate-distortion mode decision. All the above mentioned methods need modification on the current image/video coding standard, which limits their practicability. Compared with these methods, the proposed scheme is standard compliant. We envision that our scheme can be a useful enhancer of any existing video compression standard, by just adding pre- and post-processing modules, to improve low bit-rates compression performance.

3. The framework of model-based low bit-rate video coding scheme

3.1. Encoder

The major concern of our system design is to provide a light-duty encoder under energy consumption and bandwidth constraints, meanwhile, have the ability to mitigate the problem of packet losses during

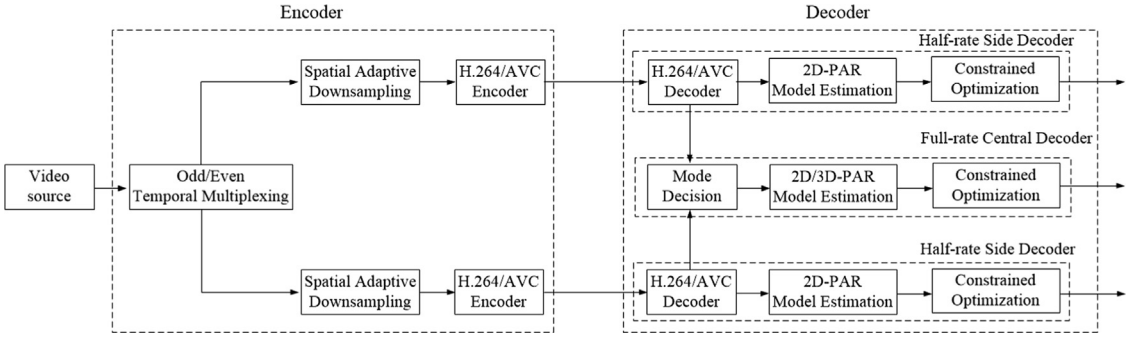

Fig. 1. The framework of the proposed video coding scheme.

wireless transmission. In this paper, we propose a new encoder scheme called temporal multiplexing with spatial adaptive downsampling, as illustrated in Fig. 1. To improve error resilience of the bitstream, we propose to divide the input video source into two side descriptions according to their spatial and temporal relationships. Specially, the input video sequence is first prefiltered with a Gaussian filter in temporal domain. The filtered video frames are then split by a simple multiplexer into odd and even side descriptions, each of which contains odd and even frames respectively. After that, each frame in two descriptions is performed spatially downsampling to further compact the video signals. The generated two low-resolution (LR) descriptions are mutually refinable and can be independently decoded.

For spatial downsampling, we choose not to perform uneven irregular down sampling of a video frame according to local spatial or frequency characteristics. Instead, we stick to conventional square pixel grid by uniform spatial down sampling of the frame with a factor of two. Yet the simple uniform downsampling scheme is made adaptive by a directional low-pass prefiltering step prior to downsampling. In this way, the two side descriptions can be compressed by any third-party encoder and transmitted to the decoder by the same or separate wireless channel.

The other purpose of this preprocessing is to induce a mechanism of collaboration between the spatial uniform downsampling process at the encoder and optimal upconversion process at the decoder. The sampling locations of odd and even descriptions are different, which are carefully designed to make the resulting LR side descriptions interleave with each other on HR sample grid in the temporal domain, as illustrated in Fig. 2. In this way, we can introduce structure redundancy for two descriptions. This design has advantages to be self-evident in the following development.

Now we can see that spatial adaptive downsampling can efficiently ease the burden of the encoder and wireless transmission channel because it can greatly reduce the amount of data needed to processing. In this way, it naturally extends the lifetime of the WVSN. Meanwhile, since a uniform downsampling scheme is chosen, the downsampled descriptions will constitute two LR video sequences, which can be compressed to further reduce the data rate by using standard encoder (e.g., H.264/AVC). Therefore, the whole system remains standard compliant and practical.

3.2. Decoder

As we have seen in the previous subsection, the proposed encoder generates two equal importance descriptions, which are individually packetized and sent through either the same or separate physical channels. Either channel may fail with probability $p_i, i=1,2$.

There are two common environments for MDC. One is the on–off setting, in which a typical assumption is that both channels do not simultaneously fail, and if a channel fails the corresponding description is totally lost [15]. This case is focused on by most of the existing MDC schemes. The other environment is packet lossy network, in which packet losses occur in each description, and both descriptions can be used at the decoder. In this case, the performance of error concealment plays a very important role on the final coding efficiency. In this paper, we focus on the on–off case. And the proposed scheme can be easily extended to the case of packet lossy network by using existing error concealment technologies to first conceal the errors in each description then use the proposed algorithm for reconstruction.

At the decoder, three situations are possible: both descriptions are received or either one of the two descriptions is received. The central decoder receives both descriptions and produces a high-quality reconstruction (full frame rate), while the two side decoders each receives only one of the two descriptions and produces lower, but still acceptable, quality reconstructions (half frame rate). In the following, we will detail the central decoder design. Note that the side decoder is performed in the same way as the Intra mode of the central decoder.

4. Mode dependent soft decoding by model-based estimation

In this section, we introduce the central decoder of the proposed video compression system. At the central decoder, the two side descriptions are individually decoded and mutually refined, which we call hard-decoded videos. In contrast, the restored high-resolution video sequence is called soft-decoded video, and the restoration process is called soft decoding. Soft decoding can be expected to improve the fidelity of hard-decoded videos because there is strong spatio-temporal correlation between two side descriptions. The side decoder is algorithmically similar to the central decoder, but uses only one side description that is successfully received at the decoder for reconstruction.

4.1. The interpolation model

The problem of soft decoding could be defined as follows: Let $\mathbf{y}$ be the low-pass filtered, down-sampled and compressed frame. The vector $\mathbf{y} \in \mathcal{Z}^M$ consists of M LR pixel values in a given lexicographical order, where $\mathcal{Z}$ is an integer alphabet from which the pixel values are drawn. What we want to do is to recover the underlying HR frame $\mathbf{x} \in \mathcal{Z}^N$. The formation of $\mathbf{y}$ from $\mathbf{x}$ is modeled as

$$\mathbf{y} = \mathbf{DHx} + \mathbf{n}, \tag{1}$$

where $\mathbf{H}$ is the low-pass filtering operation and $\mathbf{D}$ is the downsampling process. The term $\mathbf{n}$ is the quantization noise in compression. In what follows we develop a model-based reconstruction approach to perform up-sampling, inverse filtering and denoising jointly.

Reconstruction of $\mathbf{x}$ from $\mathbf{y}$ is inherently an ill-posed inverse problem. The performance of the reconstruction algorithm largely depends on how well it can employ regularization conditions or constraints when numerically solving the problem. The solution

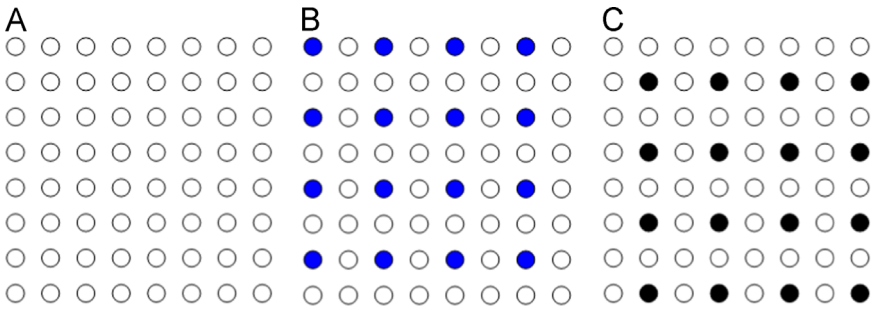

Fig. 2. Uniform downsampling with temporal multiplexing. (a) The original frame. (b) Downsampled version (blue points) for odd frames. (c) Downsampled version (black points) for even frames. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

can be greatly improved if a good adaptive video model can be integrated into the estimation process, because the model can regulate estimated pixels according to useful prior statistical knowledge.

Motivated by the geometric constraint of edges and motion trajectory, we propose to use three-dimensional piecewise auto-regressive (3D-PAR) model for video interpolation at the decoder side. In the proposed 3D-PAR model, an unknown HR pixel is estimated as a linear weighted combination of its spatio-temporal neighbors. Mathematically, let the unknown HR pixel located at (i,j) in the frame t as $x(i,j,t)$, we define the 3D-PAR model as follows:

$$x(i,j,t) = \sum_{(u,v,k)\in S(i,j,t)} a_{i,j,t}^{u,v,k} x(i+u, j+v, t+k) + n(i,j,t), \tag{2}$$

where $S(i,j,t)$ is the spatio-temporal support of the 3D-PAR model; $a_{i,j,t}^{u,v,k}$ are the model parameters, and $n(i,j,t)$ is a random perturbation independent of video signal. The model parameters are locally estimated by the least squares method (which will be introduced in the next section); therefore, the proposed 3D-PAR mode is capable of being fit video signals to achieve spatio-temporal adaptation.

4.2. Mode dependent soft decoding

We integrate the 3D-PAR model into the solution of the soft decoding problem as formulated in Eq. (1). To simplify notations, from now on we use a single index to identify 2D pixel locations, and denote $x(i,t)$ as the current pixel to interpolate in the frame t. In a local window W, our task is to jointly estimate the parameters of the interpolation model and the block of HR pixels $\mathbf{x} \in W$ such that the estimated model can optimally fit the estimated $\mathbf{x}$. Now the HR frame reconstruction from a compressed LR frame can be stated as the following constrained optimization problem:

$$\min_{\mathbf{x},\boldsymbol{\alpha}} \sum_{i\in W} \| x(i,t) - \sum_{(u,k)\in S(i,t)} \alpha_{u,k} x(i+u, t+k) \|^2,$$
$$\text{s.t.} \quad \| \mathbf{y} - \mathbf{D}\mathbf{H}\mathbf{x} \|^2 < \sigma_n^2(r), \tag{3}$$

where $\{\alpha_{u,k}\}$ are the parameters of PAR model to be fit to the waveform in local window W, and $\sigma_n^2(r)$ is the energy of the quantization noise of the compressed LR frame at bit rate r. Let L be the number of the LR pixels inside W, $\| \mathbf{y} - \mathbf{D}\mathbf{H}\mathbf{x} \|^2 < \sigma_n^2(r)$ corresponds to L inequality constraints.

Let us turn to choose an appropriate PAR model order d (i.e., the length of model parameters vector). On one hand, if the number of d is large, which means that there are many variables to estimate, when the number of equations between these variables is not enough, the problem of data overfitting will happen. On the other hand, if pixels in the local window W have weak temporal correlation with neighboring frames, such as W is a occlusion region, the accuracy of estimation will degrade heavily due to the introduction of uncorrelated temporal neighbors. Therefore, spatial PAR model (i.e., 2D-PAR) and spatio-temporal PAR model (i.e., 3D-PAR) should be adaptively chosen according to the spatio-temporal statistics of W.

On a second reflection, fortunately, the two dimensions of the image signal afford us ways to circumvent the problem of data overfitting. One way to increase the number of equations or constraints on pixels in W is the use of multiple PAR models that associate pixels in different directions. Specifically, we introduce two 6-order 3D-PAR models in our design, as illustrated in Fig. 3. One is the diagonal model $AR_\times$ which consists of four 8-connected spatial neighbors and two temporal neighbors from the forward and backward reference frames, and the other is the axial model AR_+ which consists of four 4-connected spatial neighbors and two temporal neighbors from the forward and backward reference frames. Moreover, according to the statistical duality between LR frame and its HR counterpart, the predictive mode generated in LR descriptions compression could provide us the prior knowledge to decide whether only spatial or both spatial and temporal correlation is used for HR frames reconstruction. In accordance with the H.264/AVC standard, the proposed soft-decoding reconstruction can be divided by the four modes: Intra, Skip, Inter-P, Inter-B. On a second reflection, fortunately, the two dimensions of the image signal afford us ways to circumvent the problem of data overfitting. One way to increase the number of equations or constraints on pixels in W is to use multiple PAR models that associate pixels in different directions. Specifically, we introduce two 6-order 3D-PAR models in our design, as illustrated in Fig. 3. One is the diagonal model $AR_\times$ which consists of four 8-connected spatial neighbors and two temporal neighbors from the forward and backward reference frames, and the other is the axial model AR_+ which consists of four 4-connected spatial neighbors and two temporal neighbors from the forward and backward reference frames. Moreover, according to the statistical duality between LR frame and its HR counterpart, the predictive mode generated in LR descriptions compression could provide us the prior knowledge to decide whether only spatial or both spatial and temporal correlation is used for HR frames reconstruction. In accordance with the H.264/AVC standard, the proposed soft-decoding reconstruction can be divided by four modes: Intra, Skip, Inter-P, Inter-B.

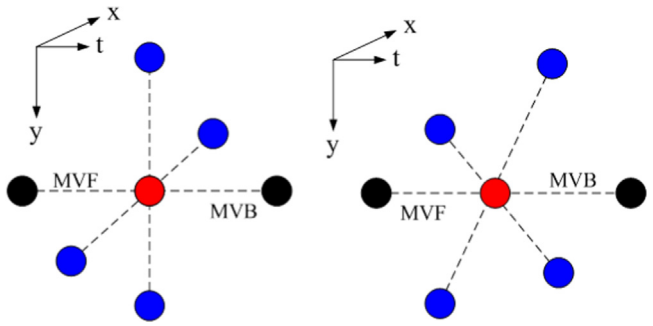

Fig. 3. Two used 6-order 3D-PAR model. The central red pixel is the one to estimate. The blue pixels are spatial neighbors. The black pixels are temporal neighbors, which are aligned by forward and backward MVs. The left model is the diagonal model $AR_\times$, which consists of four 8-connected spatial neighbors and two temporal neighbors from the forward and backward reference frames. The right one is the axial model AR_+, which consists of four 4-connected spatial neighbors and two temporal neighbors from the forward and backward reference frames. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

4.2.1. Intra mode

For Intra mode, the problem of frame reconstruction degrades to spatial image interpolation. The upconversion is based on the diagonal and axial 2D-PAR image models and on the deconvolution of the directional low-pass prefiltering. Incorporating these two PAR models into the original nonlinear estimation framework, we state the task of upconversion as the following constrained least squares problem:

$$\min_{\mathbf{x},\mathbf{a},\mathbf{b}} \left\{ \begin{array}{l} \zeta^{\times} \sum\limits_{i\in W} \| x(i,t) - \mathbf{a}_s^T \mathbf{s}_\times(i,t) \|^2 \\ + \zeta^{+} \sum\limits_{i\in W} \| x(i,t) - \mathbf{b}_s^T \mathbf{s}_+(i,t) \|^2 + \lambda \| \mathbf{y} - \mathbf{D}\mathbf{H}\mathbf{x} \|^2 \end{array} \right\}, \tag{4}$$

where $\mathbf{s}_\times(i,t)$ and $\mathbf{s}_+(i,t)$ consist of four 8-connected and four 4-connected spatial neighbors of $x(i,t)$ in the HR image, $\mathbf{a}_s$ and $\mathbf{b}_s$ are model parameters of diagonal and axial models, respectively, $\zeta^{\times}$ and ζ^{+} are fusion weights to combine the modeling strength of the two PAR models.

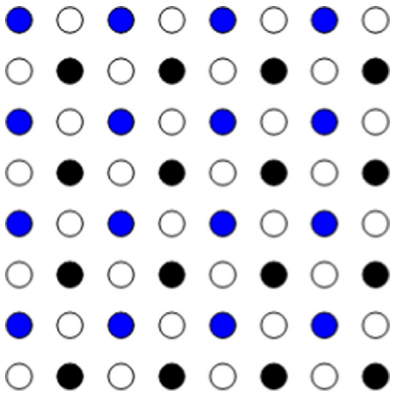

Fig. 4. Quincunx sample grid. The blue dots are LR samples in the current frame, the black dots are copied from the forward frame, the blank dots are HR samples needed to estimate. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

4.2.2. Skip mode

For Skip mode, the LR pixels in the forward frame $t-1$ can be directly copied to the current sample grid to construct a quincunx lattice, as illustrated in Fig. 4. With the quincunx lattice, the Skip mode performs spatial interpolation to estimate other missing pixels. The optimization formulation is the same as Eq. (4), while with $2L$ inequality constraints since two times LR samples can be available. The increased constraints can provide more accurate estimation from the solution space.

4.2.3. Inter-P and inter-B mode

For Inter-P and Inter-B mode, motion information is available to facilitate the task of resolving intensity uncertainty of video signals by exploiting the fundamental tradeoff between spatial and temporal correlation. Two 3D-PAR models are used for such case, where the current pixel is approximated as the weighted combination of samples within its spatial neighborhoods as well as the temporal neighbors aligned by motion vectors. The task of upconversion can be stated as the following constrained least squares problem:

$$\min_{\mathbf{x},\mathbf{a},\mathbf{b}} \left\{ \begin{array}{l} \zeta^{\times} \sum\limits_{i\in W} \| x(i,t) - (\mathbf{a}_s^T \mathbf{s}_{\times}(i,t) + \mathbf{a}_t^T(i,t)) \|^2 \\ +\zeta^{+} \sum\limits_{i\in W} \| x(i,t) - (\mathbf{b}_s^T \mathbf{s}_{+}(i,t) + \mathbf{b}_t^T(i,t)) \|^2 \\ +\lambda \| \mathbf{y} - \mathbf{D}\mathbf{H}\mathbf{x} \|^2 \end{array} \right\}, \tag{5}$$

where $\mathbf{a}_s$ and $\mathbf{b}_s$ are spatial model parameters along diagonal and axial direction, respectively; $\mathbf{a}_t$ and $\mathbf{b}_t$ are temporal model parameters along the motion vector; (i,t) is the temporal reference sample set which includes forward reference sample for the Inter-P mode and bi-directional reference samples for the Inter-B mode.

5. Model estimation and convex optimization

The PAR model, via its parameters, offers an adaptive sparse representation of video signals. Therefore, 2D-PAR and 3D-PAR model estimations are a critical issue in the proposed low bit-rate video coding scheme. In the following, we will show in detail how to derive the model parameters. Besides, we will show how the process of video reconstruction can be formulated as a convex optimization problem and finally derive a closed-form solution.

5.1. Model parameters estimation

The model parameters specify the direction and amplitude of edges in spatial and motion in temporal. They are estimated on the fly for each pixel using sample statistics of a local spatio-temporal covering. It accounts for the fact that the spatio-temporal statistics of natural video signals are often piecewise stationary. The accuracy of model parameter estimation directly influences the quality of reconstructed frames. In the following, let us consider how to estimate the model parameters $\mathbf{a}_s$ and $\mathbf{b}_s$ for 2D-PAR model used in Intra and Skip mode, and $\mathbf{a}=[\mathbf{a}_s,\mathbf{a}_t]$ and $\mathbf{b}=[\mathbf{b}_s,\mathbf{b}_t]$ for 3D-PAR mode used in Inter-P and Inter-B mode.

For 2D-PAR model, our emphasis is to reconstruct significant edges. The study of natural image statistics reveals that the second order statistics of natural images tends to be invariant across different scales, and those scale invariant features are shown to be crucial for visual perception [21,22]. Therefore, we can estimate local covariance coefficients from a low-resolution image, and then project the estimated covariance to the high-resolution image to adapt the interpolation. Moveover, in order to reduce the influence of compression noise and enhance the robustness of estimated model, we propose to learn model parameters $\mathbf{a}_s$ and $\mathbf{b}_s$ from decoded image using moving least square. Specifically, for a local window W centered on $\mathbf{x}(i,t)$, we estimate the model parameters by solving the following two optimization problems:

$$\begin{aligned} \mathbf{a}_s^* &= \min_{\mathbf{a}_s} \left\{ \sum_{j\in W} \theta(i,j) \| y(j,t) - \mathbf{a}_s^T \mathbf{s}_y^{\times}(j,t) \|^2 + \lambda \| \mathbf{a}_s \|^2 \right\}, \\ \mathbf{b}_s^* &= \min_{\mathbf{b}_s} \left\{ \sum_{j\in W} \theta(i,j) \| y(j,t) - \mathbf{b}_s^T \mathbf{s}_y^{+}(j,t) \|^2 + \lambda \| \mathbf{b}_s \|^2 \right\}, \end{aligned} \tag{6}$$

where $\mathbf{s}_y^{\times}$ and $\mathbf{s}_y^{+}$ are samples along diagonal and axial direction in the LR frame $\mathbf{y}$, respectively, $\theta(i,j)$ is the moving weight to reflect the similarity of the sample on j with the sample on the center i. In this paper, we combine the edge-preserving property of bilateral filter and the robust property of non-local-means weights to design effective moving weights, which is defined as follows:

$$\begin{aligned} &\theta(i,j) = \frac{1}{N} \exp\left\{ -\frac{\| i-j \|^2}{\sigma_s^2} \right\} \exp\left\{ -\frac{G \cdot \| SW(i) - SW(j) \|^2}{\sigma_p^2} \right\}, \\ &\sigma_s > 0, \sigma_q > 0. \end{aligned} \tag{7}$$

where N is the normalization factor, G is a Gaussian kernel used to take into account the distance between the central pixel and other pixels in the patch, and $SW(i)$ represents the pixel patch whose components are intensity values of pixels in the similarity window centered on i.

Setting the derivative to 0, the closed form solutions of Eq. (6) are

$$\begin{aligned} \mathbf{a}_s^* &= (\mathbf{A}^T \mathbf{\Theta} \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^T \mathbf{\Theta} \mathbf{v}; \\ \mathbf{b}_s^* &= (\mathbf{B}^T \mathbf{\Theta} \mathbf{B} + \lambda \mathbf{I})^{-1} \mathbf{B}^T \mathbf{\Theta} \mathbf{v}, \end{aligned} \tag{8}$$

where the column vector $\mathbf{v}$ is composed of all $y(i,t)$ in W. The ith row of the matrix $\mathbf{A}$ consists of the four 8-connected neighbors of $y(i,t)$, and the ith row of the matrix $\mathbf{B}$ consists of the four 4-connected neighbors of $y(i,t)$. $\mathbf{\Theta}$ is a diagonal matrix whose entries in diagonal locations are moving weights.

As formulated in Eq. (4), we use $\zeta^{\times}$ and ζ^{+} to combine the modeling strength of the two PAR models. We can exploit the squared errors associated with the solutions of two objective functions in Eq. (8) to determine these two fusion weights:

$$\begin{aligned} \zeta^{\times} &= \frac{e^{+}}{e^{+}+e^{\times}}, \\ \zeta^{+} &= \frac{e^{\times}}{e^{+}+e^{\times}}. \end{aligned} \tag{9}$$

These weights are optimal in least squares sense if the fit errors of the two models are independent.

In still images, intensity field is homogeneous along the edge orientation. If considering the counterpart of edge contours in 3D, we observe that intensity field is homogeneous along the motion

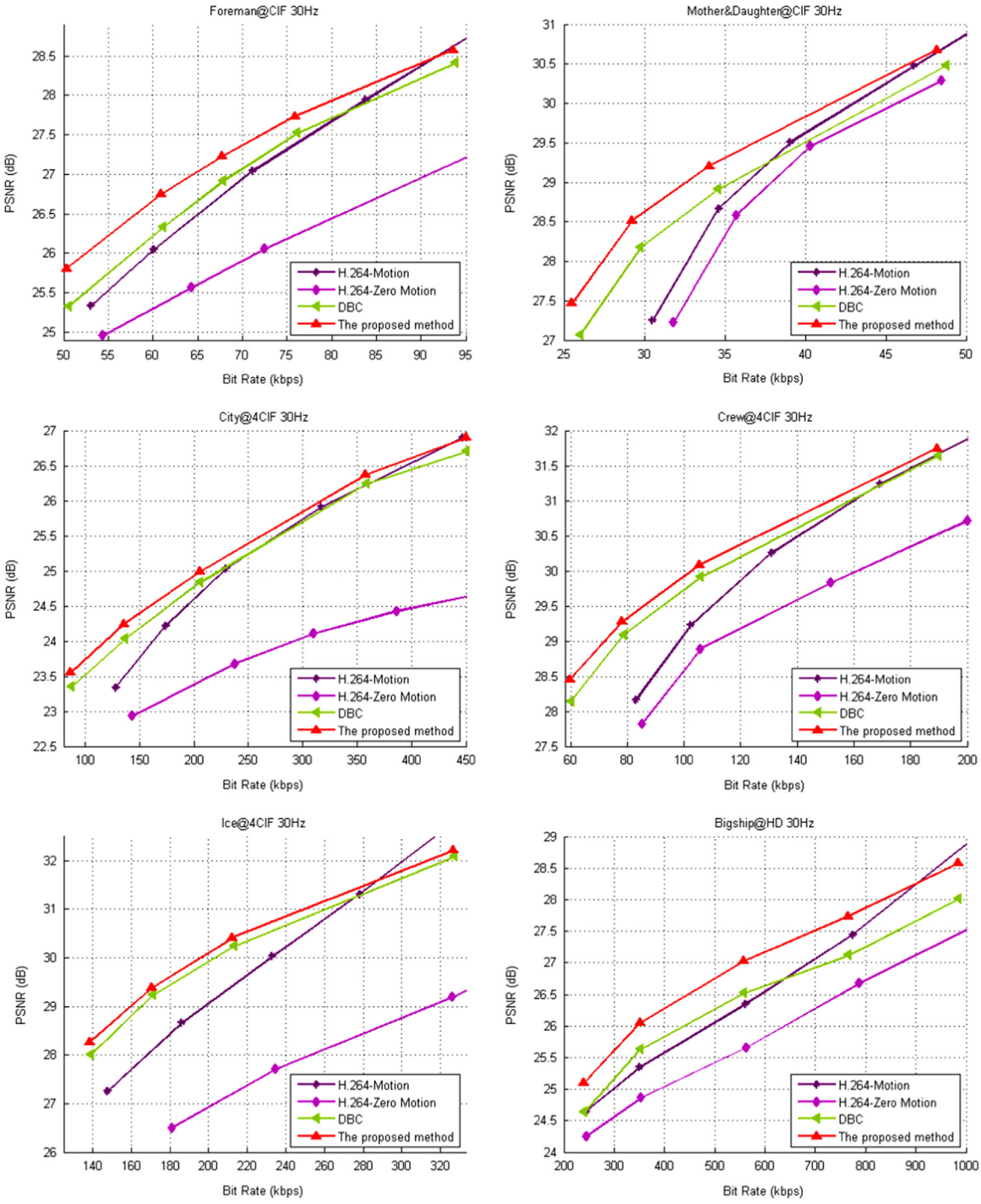

Fig. 5. Rate–distortion performance comparison.

trajectory [24]. The duality between edge contour and motion trajectory motivates us to use the similar approach as that of 2D-PAR for model estimation of 3D-PAR. For 3D-PAR model, the model parameters are adaptively estimated within a localized spatio-temporal window from the side description video sequence by the moving least-square method [26] on a pixel-by-pixel basis. Similarly, the derivation $\mathbf{a}=[\mathbf{a}_s,\mathbf{a}_t]$ and $\mathbf{b}=[\mathbf{b}_s,\mathbf{b}_t]$ follows the moving least squares formulation:

$$\mathbf{a}^*=\min_{\mathbf{a}}\left\{\begin{array}{l}\sum\limits_{j\in W}\theta(i,j)\|y(j,t)-(\mathbf{a}_s^T\mathbf{s}_y^{\times}(j,t)+\mathbf{a}_t^T\mathbf{t}_y(j,t))\|^2\\ \quad+\lambda\|\mathbf{a}\|^2\end{array}\right\},$$

$$\mathbf{b}^*=\min_{\mathbf{b}}\left\{\begin{array}{l}\sum\limits_{j\in W}\theta(i,j)\|y(j,t)-(\mathbf{b}_s^T\mathbf{s}_y^{+}(j,t)+\mathbf{b}_t^T\mathbf{t}_y(j,t))\|^2\\ \quad+\lambda\|\mathbf{b}\|^2\end{array}\right\},\tag{10}$$

with the closed-form solutions:

$$\mathbf{a}^*=(\widehat{\mathbf{A}}^T\mathbf{\Theta}\widehat{\mathbf{A}}+\lambda\mathbf{I})^{-1}\widehat{\mathbf{A}}^T\mathbf{\Theta}\mathbf{v};$$

$$\mathbf{b}^*=(\widehat{\mathbf{B}}^T\mathbf{\Theta}\widehat{\mathbf{B}}+\lambda\mathbf{I})^{-1}\widehat{\mathbf{B}}^T\mathbf{\Theta}\mathbf{v},\tag{11}$$

where the ith row of matrix $\widehat{\mathbf{A}}$ consists of the four 8-connected spatial neighbors, and one or two temporal neighbors of $y(i,t)$; the ith row of matrix $\widehat{\mathbf{B}}$ consists of the four 4-connected spatial

neighbors and one or two temporal neighbors of $y(i,t)$. We can obtain $\zeta^{\times}$ and ζ^{+} in a similar way with Eq. (9).

5.2. Convex optimization

Once the PAR model is constructed, soft decoding can be performed efficiently by constrained linear least squares estimation. Considering that the prediction of the 2D-PAR or 3D-PAR model is linear combinations of the pixels in a spatial or spatio-temporal neighborhood, we can arrange the estimated parameters $\{\mathbf{a}\}$ and $\{\mathbf{b}\}$ of the diagonal and axial model into two sparse matrixes $\mathbf{A}$ and $\mathbf{B}$, and finally incorporate the model into the objective function. For convenient representation we rewrite Eqs. (4) and (5) in the matrix form:

$$\mathbf{x}^{*}=\min_{\mathbf{x}}\left\{\begin{array}{l}\zeta^{\times}\sum\limits_{i\in W}\|\mathbf{x}-\mathbf{A}\mathbf{x}\|^{2}+\zeta^{+}\sum\limits_{i\in W}\|\mathbf{x}-\mathbf{B}\mathbf{x}\|^{2}\\ +\lambda\|\mathbf{y}-\mathbf{D}\mathbf{H}\mathbf{x}\|^{2}\end{array}\right\}. \tag{12}$$

The objective function can be further written in quadratic form as

$$\min_{\mathbf{x}}\mathbf{r}(\mathbf{x})^{T}\mathbf{r}(\mathbf{x}), \tag{13}$$

where the residue vector $\mathbf{r}(\mathbf{x})$ is defined as

$$\mathbf{r}(\mathbf{x})=\begin{bmatrix}\sqrt{\zeta^{\times}}(\mathbf{I}-\mathbf{A})\mathbf{x}\\ \sqrt{\zeta^{+}}(\mathbf{I}-\mathbf{B})\mathbf{x}\\ \sqrt{\lambda}(\mathbf{y}-\mathbf{D}\mathbf{H}\mathbf{x})\end{bmatrix}. \tag{14}$$

And the objective function in Eq. (13) is a linear least square problem that can obtain a closed-form solution as

$$\mathbf{x}=(\mathcal{F}^{T}\mathcal{F})^{-1}\mathcal{F}^{T}\mathcal{G} \tag{15}$$

with

$$\mathcal{F}=\begin{bmatrix}\sqrt{\zeta^{\times}}(\mathbf{I}-\mathbf{A})\\ \sqrt{\zeta^{+}}(\mathbf{I}-\mathbf{B})\\ \sqrt{\lambda}\mathbf{D}\mathbf{H}\end{bmatrix} \tag{16}$$

and

$$\mathcal{G}=\begin{bmatrix}0\\ 0\\ -\sqrt{\lambda}\mathbf{y}\end{bmatrix}. \tag{17}$$

6. Experimental results

In this section, experimental results are presented to verify the performance of the proposed video coding scheme with respect to rate-distortion performance, subjective quality and encoder complexity. For thoroughness and fairness of our comparison study, we selected six video sequences as test ones, including two CIF sequence: *Foreman*, *Mother and Daughter*; three 4CIF sequence: *City*, *Crew*, *Ice*; and one HD sequence: *Bigship*. All of them are with frame rate 30 Hz, and each sequence contains 100 frames.

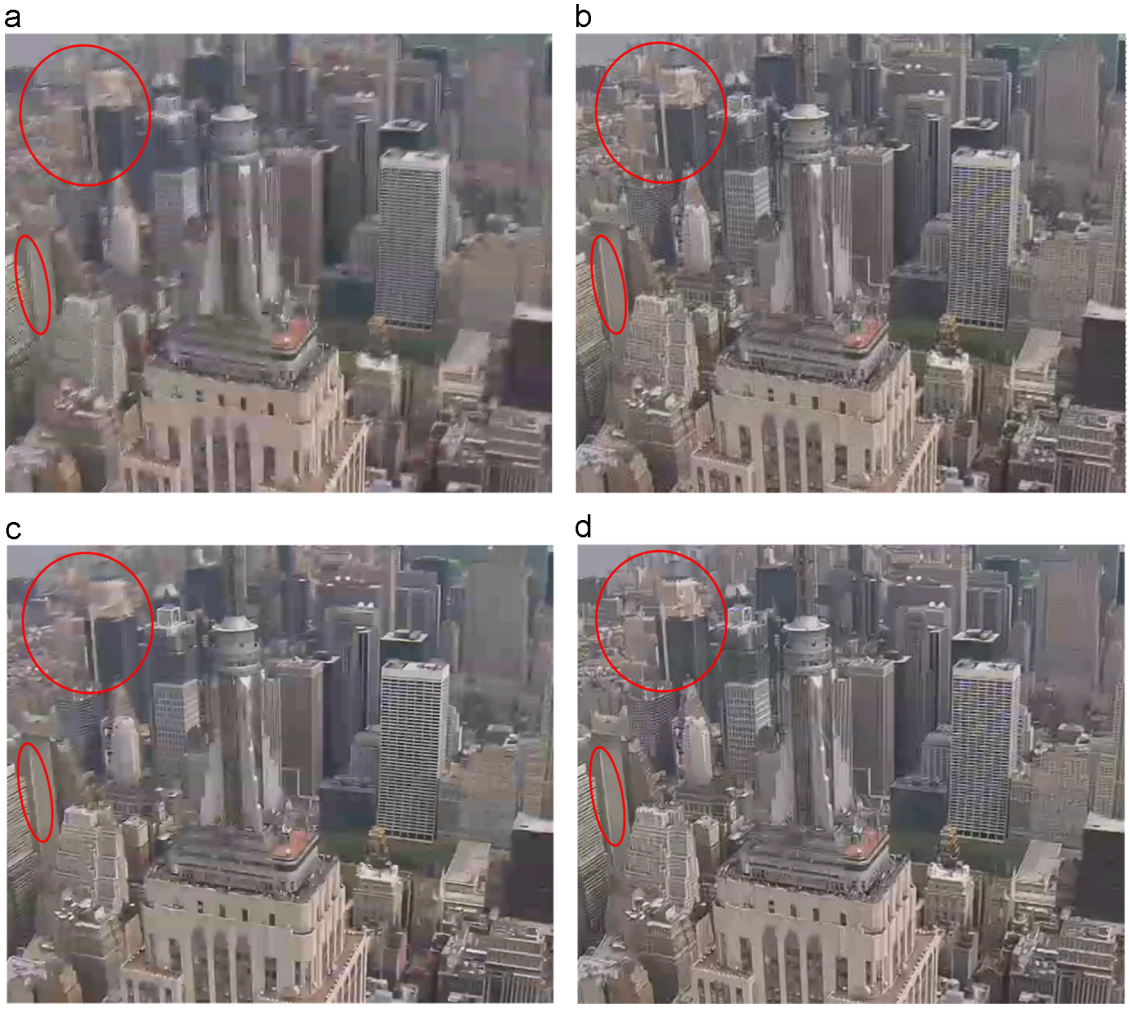

Fig. 6. Subjective quality comparison of reconstructed second frame in *City* sequence with (PSNR, SSIM, Bit Rate). (a) H.264 Zero Motion(24.12db, 0.6013, 626.16kbps) (b) DBC, (27.01db, 0.7076, 635.54 kbps) (c) H.264-Motion (27.87db, 0.6857, 610.19 kbps) and (d) Our method (27.46db, 0.7132, 625.14 kbps).

The following video codecs will be used as benchmarks to evaluate the performance of the proposed codec.

- H.264-Motion: This codec is performed on JM16.0 [23] in main profile exploiting spatial and temporal redundancy (i.e., intra and inter prediction are both selected). The GOP size is 6 with IBPBPB structure. The RD optimization is done in the high-complexity mode. The loop filter is enabled. Entropy coding is performed in CABAC mode, and the search range of motion estimation is set to 32. It can be considered as the state-of-the-art encoder-centralized video codec.
- H.264-Zero Motion: We use JM16.0 [23] in main profile to code the GOP of 24 frames with the first frame coded as I frame and all other frames coded as predictive frames, for which the rest settings are the same as H.264-Motion except that the range of motion search is set to 1. It is often used as a benchmark in comparison for non-ME based low complexity video codec.
- DBC: This is a state-of-the-art downsampling-based video coding scheme [18].

Since DVC and H.264-Intra give too poor RD performances, we have not included their results into comparison. The performance comparison among our method and them can be indirectly reflected by H.264/AVC standard codecs. Through comparisons with these two standard codecs motioned above, results are shown that serve to support the efficiency of the proposed scheme for low bit-rate coding.

6.1. Low bit-rate performance comparison

To verify the performance improvement of the proposed scheme at low bit rates, we use coarse quantization parameters (QP) to obtain rate–distortion curves shown below. And the comparisons shown here are all for approximately the same average bit rate over the entire sequence and therefore can be readily compared in terms of the PSNR values. The RD curves of six

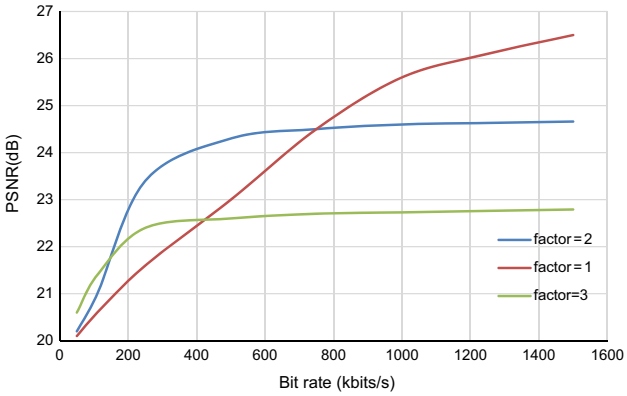

Fig. 8. Analytical rate distortion curves of different downsampling factors.

Fig. 7. Subjective quality comparison of the reconstructed second frame in *Ice* sequence with (PSNR, SSIM, Bit Rate). (a) H.264 Zero Motion(28.68db, 0.8703, 397.27kbps) (b) DBC, (33.01db, 0.9203, 423.76 kbps) (c) H.264-Motion (33.77db, 0.9073, 401.01 kbps) and (d) Our method (33.36db, 0.9259, 413.44 kbps).

video sequences are plotted in Fig. 5. We can find for most of the six test sequences the proposed scheme can achieve better RD performance at low bit-rate compared with other five codecs. The gain is up to 1.5 dB for CIF sequences and 1.2 dB for 4CIF sequences compared with H.264-Motion, which is regarded as the state-of-the-art video codec. Compared with H.264-Zero Motion, which is also with a low-complexity encoder, the gains are obvious and is up to 1.5 dB for CIF sequences and 3 dB for 4CIF sequences. According to the trend of RD curves, the proposed method outperforms H.264-Zero Motion at a wide range of bit-rate.

Our method outperforms DBC as well. In DBC, the training image is with original size and compressed by intra mode at low bit-rate; therefore, its quality is poor. The quantization noise within it will propagate to other frames by the process of SR. Moreover, if there is irregular motion in the GOP, such as the *City* sequence, the training image fails to provide good prior information.

6.2. Medium bit-rate performance comparison

The advantage of the proposed scheme is not only limited to low bit-rates but also gives the subjective comparison results at medium bit-rates. As illustrated in Figs. 6 and 7, we show the decoded frames of two 4CIF sequences by the proposed side and central decoders and other compared codecs. From the results, it is easy to see that H.264-Zero Motion produces objectionable visual artifacts (e.g., jaggies and ringings) in edge areas, H.264-Motion performs better but still suffers from annoying blurring artifacts along the edges. The proposed schemes on side and central decoder are both largely free of those defects. Even when the bit rate gets higher and H.264-Motion starts to have higher PSNR than the proposed method, its visual quality still appears inferior, as demonstrated by examples in Figs. 6 and 7. This is due to the fact that quantization in H.264/AVC standard is uniform and there is no special mechanism to preserve edges that are important for human visual perception. The proposed method produces significantly enhanced perceptual quality. The results are visually compelling in reconstructing edges and textures. The produced edge and texture are clean and sharp, and most visual artifacts appeared in the results of H.264-Motion are eliminated in the proposed method. The proposed method also achieves better subjective quality compared with DBC. These results demonstrate that the proposed method can efficiently favor the reconstruction of edges. The superior visual quality of the proposed method is due to the good fit of the piecewise autoregressive model to edge structures and the fact that human visual system is highly sensitive to phase errors in reconstructed edges.

6.3. The influence of downsampling factor

Ref. [25] shows why down-sampling based image coding can beat traditional image compression standard from a perspective of rate–distortion analysis. [18] extends the analytical model to downsampling-based video coding. Since this analytical model is general and not involved to specific interpolation algorithm, here, we borrow the analytical model of [18] to show how the downsampling factors influence the final RD performance. The RD curves of different downsampling factors are shown in Fig. 8. Here, factor=1 actually represents the traditional video coding standard. As illustrated in Fig. 8, both factor=2 and factor=3 beat the traditional video coding standard at low bit-rates. By downsampling before compression, the bit budget for each DCT coefficient is increased. As such, more bits are allocated to DC and lower frequency AC coefficients, thus less quantization error is introduced and consequently better decoded quality is achieved. We note that factor=3 wins factor=1 and factor=2 only at very low bit-rates. The case factor=2 achieves good tradeoff between rate and distortion. Therefore, in this paper, we choose the downsampling factor as 2.

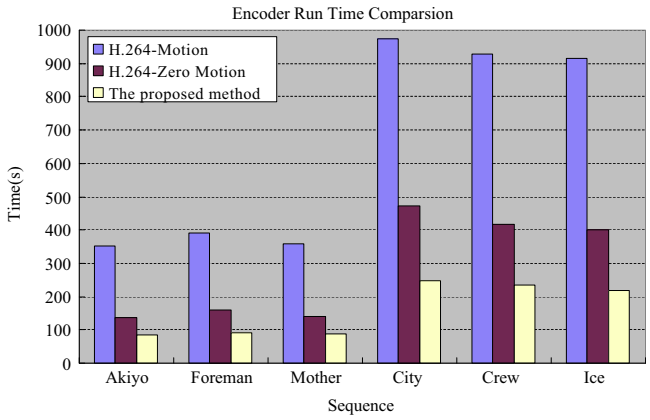

Fig. 9. Encoder complexity comparison.

6.4. Encoder complexity comparison

We also give the encoder running time comparison of the compared three codecs on six test sequences. The compared codecs are run on a typical computer (2.5 GHz Intel Dual Core, 4G Memory). For each sequence, we keep the bit-rates of three codecs almost the same. As illustrated in Fig. 9, it is easy to find that the proposed method achieves lowest encoder complexity, the running time is even lower than H.264-Zero Motion. The running time is about 1/4 that of H.264-Motion, because the video data needed to be compressed reduces to 1/4 of original ones by downsampling. These results demonstrate our method can provide a lightweight encoder, which is attractive for resource-deficient wireless video communications.

7. Conclusion

In this paper, we presented an effective and flexible low bit-rate video coding scheme through combing the idea of sparse sampling and multiple description coding for wireless video communications. At the encoder, multiple low-resolution descriptions are generated by temporal multiplexing and spatial adaptive downsampling. Based on the PAR image model and predictive modes in compression, the decoder solves an inverse problem to jointly estimate the model coefficients and the high resolution frame. Experimental results demonstrate that the proposed video coding scheme outperforms H.264/AVC and other state-of-the-art methods at low bit-rates.

Acknowledgments

This work was supported by the Major State Basic Research Development Program of China (973 Program 2015CB351804), the National Science Foundation of China under Grant nos. 61300110, 61272386, 61371146 and 61402547, the Natural Scientific Research Innovation Foundation of HIT under Grant nos. KMQQ5750010315 and KMQQ5750009614, and the Macau Science and Technology Development Fund under grant nos. FDCT/009/2013/A1, FDCT/046/2014/A1.

References

[1] I.F. Akyildiz, T. Melodia, K.R. Chowdhury, Wireless multimedia sensor networks: a survey, IEEE Wirel. Commun. Mag. 14 (December (6)) (2007) 32–39.
[2] I.F. Akyildiz, T. Melodia, K.R. Chowdhury, A survey on wireless multimedia sensor networks, Comput. Netw. 51 (March (4)) (2007) 921–960.
[3] S. Soro, W. Heinzelman, A survey of visual sensor networks, Adv. Multimed. 2009 (2009) (Article ID 640386).
[4] G. Anastasi, M. Conti, M. di Francesco, A. Passarella, Energy conservation in wireless sensor networks: a survey, Ad Hoc Netw. 7 (May (3)) (2009) 537–568, Elsevier.
[5] Z. He, Y. Liang, L. Chen, I. Ahmad, D. Wu, Power-rate-distortion analysis for wireless video communication under energy constraint, IEEE Trans. Circuits Syst. Video Technol. 15 (May (5)) (2005) 645–658.
[6] B. Haskell, A. Puri, A. Netravali, Digital Video: An Introduction to MPEG-2, Chapman and Hall, New York, 1996.
[7] T. Wiegand, G. Sullivan, G. Bjontegaard, A. Luthra, Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol. 13 (July (7)) (2003) 560–576.
[8] B. Girod, A. Aaron, S. Rane, D. Rebollo-Monedero, Distributed video coding, Proc. IEEE 93 (January (1)) (2005) 1–12.
[9] R. Puri, A. Majurndar, K. Ramchandran, PRISM: a video coding paradigm with motion estimation at the decoder, IEEE Trans. Image Process. 16 (October (10)) (2007) 1–13.
[10] Z. Xiong, A.D. Liveris, S. Cheng, Distributed source coding for sensor networks, IEEE Signal Process. Mag. 21 (5) (2004) 80–94.
[11] A. Aaron, E. Setton, B. Girod, Towards practical Wyner-Ziv coding of video, in: Proceedings of IEEE International Conference on Image Processing (ICIP), Barcelona, Spain, September 2003.
[12] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, M. Quaret, The DISCOVER codec: architecture, techniques and evaluation, in: Picture Coding Symposium, Lisboa, Portugal, November 2007.
[13] X. Liu, D. Zhao, Y. Zhang, S. Ma, Q. Huang, W. Gao, Joint learning for side information and correlation model based on linear regression model in distributed video coding, in: Proceedings of IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, November 7–10, 2009.
[14] V.K. Goyal, Multiple description coding: compression meets the network, IEEE Signal Process. Mag. 18 (September) (2001) 74–93.
[15] Y. Wang, A. Reibman, S. Lin, Multiple description coding for video delivery, Proc. IEEE 93 (2005) 57–70.
[16] T. Tillo, E. Baccaglini, G. Olmo, Multiple descriptions based on multirate coding for JPEG 2000 and H.264/AVC, IEEE Trans. Image Process. 19 (July (7)) (2010) 1756–1767.
[17] D. Wang, N. Canagarajah, D. Bull, Slice group based multiple description video coding using motion vector estimation, in: Proceedings of IEEE International Conference on Image Processing (ICIP), September 2004.
[18] M. Shen, P. Xue, C. Wang, Down-sampling based video coding using super-resolution technique, IEEE Trans. Circuits Syst. Video Technol. 21 (6) (2011) 755–765.
[19] M. Biswas, M.R. Frater, J.F. Arnold, Multiple description wavelet video coding employing a new tree structure, IEEE Trans. Circuits Syst. Video Technol. 18 (October (10)) (2008) 1361–1368.
[20] V.A. Nguyen, Y.P. Tan, W.S. Lin, Adaptive downsampling/upsampling for better video compression at low bit rate, in: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2008, pp. 1624–1627.
[21] A. Srivastava, A.B. Lee, E.P. Simoncelli, S.-C. Zhu, On advances in statistical modeling of natural images, J. Math. Image Vis. 18 (1) (2003) 17–33.
[22] X. Liu, D. Zhai, D. Zhao, G. Zhai, W. Gao, " Progressive Image Denoising through Hybrid Graph Laplacian Regularization: A Unified Framework", IEEE Trans. on Image Process. 23 (2014).
[23] H.264/AVC reference software, JM16.0, Online available: ⟨http://iphome.hhi.de/suehring/tml/ ⟩.
[24] Xin Li, Video processing via implicit and mixture motion models, IEEE Trans. Circuits Syst. Video Technol. 17 (8) (2007) 953–963.
[25] A. Bruckstein, M. Elad, R. Kimmel, Down scaling for better transform compression, IEEE Trans. Image Process. 12 (September (9)) (2003) 1132–1144.
[26] X. Liu, D. Zhao, R. Xiong, S. Ma, W. Gao, H. Sun, "Image Interpolation via Regularized Local Linear Regression", IEEE Trans. on Image Process. 20 (2011).
[27] T. Tillo, M. Grangetto, G. Olmo, Redundant slice optimal allocation for H.264 multiple description coding, IEEE Trans. Circuits Syst. Video Technol. 18 (January (1)) (2008) 59–70.
[28] Y. Xu, C. Zhu, End-to-end rate-distortion optimized description generation for H.264 multiple description video coding, IEEE Trans. Circuits Syst. Video Technol. 23 (9) (2013) 1523–1536.
[29] S. Nazir, V. Stankovic, H. Attar, L. Stankovic, S. Cheng, Relay-assisted rateless layered multiple description video delivery, IEEE J. Sel. Areas Commun. 31 (August) (2013) 1629–1637.
[30] Y. Zhang, D. Zhao, J. Zhang, R. Xiong, W. Gao, Interpolation dependent image downsampling, IEEE Trans. Image Process. 20 (November (11)) (2011) 3291–3296.

Xianming Liu is currently an Associate Professor with the Department of Computer Science, Harbin Institute of Technology, Harbin, China. He received the B.S., M.S., and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 2006, 2008 and 2012, respectively. From 2009 to 2012, he was with National Engineering Lab for Video Technology, Peking University, Beijing, as a research assistant. In 2011, he spent half a year at the Department of Electrical and Computer Engineering, McMaster University, Canada, as a visiting student. From December 2012 to December 2013, he worked as a post-doctoral fellow at McMaster University. In 2014, he worked as Post-Doctoral Fellow with the National Institute of Informatics, Tokyo, Japan. His research interests include image/video coding, image/video processing, and machine learning.

Xinwei Gao received the B.S., and M.S. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 2008 and 2010, respectively, where he is now pursuing the Ph.D degree. His research interests include image/video coding, image/video processing.

Debin Zhao received the B.S., M.S., and Ph.D. degrees in computer science from Harbin Institute of Technology, China in 1985, 1988, and 1998, respectively. He is now a professor in the Department of Computer Science, Harbin Institute of Technology. He has published over 200 technical articles in refereed journals and conference proceedings in the areas of image and video coding, video processing, video streaming and transmission, and pattern recognition.

Jiantao Zhou is currently an Assistant Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau. He received the B. Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, Dalian, China, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, Nanjing, China, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2009. He held various research positions at the University of Illinois at Urbana-Champaign, the Hong Kong University of Science and Technology, and the McMaster University. His research interests include multimedia security and forensics, and high-fidelity image compression. He was a co-author of a paper that received the Best Paper award in the IEEE Pacific-Rim Conference on Multimedia (PCM) in 2007.

Guangtao Zhai received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2006 to 2007, he was a Student Intern with the Institute for Infocomm Research, Singapore. From 2007 to 2008, he was a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.

Wen Gao received the Ph.D. degree in electronics engineering from the University of Tokyo, Japan, in 1991. He is a professor of computer science at Peking University, China. Before joining Peking University, he was a professor of computer science at Harbin Institute of Technology from 1991 to 1995, and a professor at the Institute of Computing Technology of Chinese Academy of Sciences. He has published extensively including five books and over 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. Dr. Gao served or serves on the editorial board for several journals, such as IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, IEEE Transactions on Autonomous Mental Development, EURASIP Journal of Image Communications, Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations.