# Spatially Scalable Video Coding For HEVC

Zhongbo Shi, Xiaoyan Sun, *Senior Member, IEEE,* and Feng Wu, *Fellow, IEEE*

***Abstract*—Spatially scalable video coding (SSVC) provides an efficient way to transmit one video at different resolutions. Based on the emerging High Efficiency Video Coding (HEVC), we propose an SSVC scheme to support both single-loop (SL) and multiloop (ML) solutions by enabling different interlayer prediction mechanisms. Specifically, we employ two interlayer prediction modes: quadtree-based prediction mode (Q-mode) and learning-based prediction mode (L-mode). The Q-mode is investigated to exploit the interlayer redundancy based on the quadtree coding structure of HEVC. Due to the high correlation between layers, Q-mode utilizes the coded information from the base layer quadtree, including coding unit split, prediction unit partition, motion information, and partial texture information of transform unit, to predict the enhancement layer quadtree. By enabling Q-mode, we provide a basic SL solution for low complexity applications. Besides the correlation explored in Q-mode, we employ an extra L-mode to further improve the coding performance. In L-mode, the temporal-spatial correlation is exploited simultaneously by visual patch-based learning and mapping at pixel level. This helps us achieve more accurate prediction signals based on the coarse base layer reconstruction within an ML structure. Experimental results show the effectiveness of our SSVC scheme compared with the simulcast case and other HEVC-based SSVC schemes.**

***Index Terms*—High Efficiency Video Coding (HEVC), learning-based approach, scalable video coding (SVC).**

## I. Introduction

DURING THE last decade, traditional server–client video streaming has been unable to meet people's ever growing demands for video applications. With the boom of mobile devices, more and more people share and browse video content on social networks via smart phones or tablets. Broadband networks, especially 3G/4G wireless networks, make video communication a part of people's daily lives. In these new application scenarios, different kinds of devices are involved in a big system as shown in Fig. 1, in which different screen resolutions, computing capabilities and demands for different network bandwidths among these devices create a real problem—how to meet these various needs efficiently. Traditional video coding schemes do not work well. Scalable video coding (SVC) presents a highly attractive solution to these applications, since it enables a single bitstream to simultaneously serve various devices with different display resolutions and qualities. Furthermore, the rapid developments on multimedia DSP chips support the implementation of a real SVC system. Hence, the investigation on SVC, especially spatially SVC (SSVC), is necessary and important.

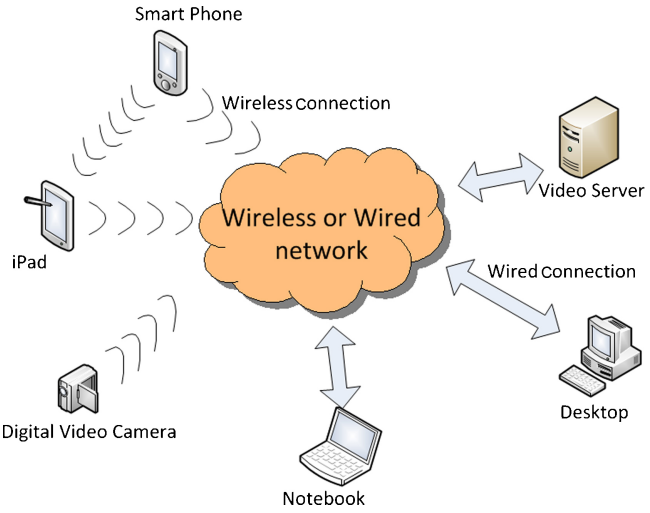



Fig. 1. Different devices in one video system.

SVC has been researched for more than 20 years. Most inter national video coding standards, such as H.262|MPEG-2 Video [1], H.263 [2], MPEG-4 [3], support various scalabilities to meet various needs for distributing videos in heterogeneous environments. However, due to the decreased coding efficiency as well as the high complexity, those SVC solutions have limited applications. The latest SVC standard—the scalable extension of H.264/Advanced Video Coding (AVC) [4], [5], employs a so-called single-loop (SL) motion compensation framework to better balance the coding efficiency and the decoding complexity than prior standards. But the gap between SVC schemes and non-SVC ones in terms of coding efficiency is still wide. Recently, MPEG has been developing the next generation video coding standard—High Efficiency Video Coding (HEVC) [6]. New and state-of-the-art techniques and coding tools are being adopted to significantly improve coding performance compared with H.264/AVC. Currently, MPEG is calling for proposals on scalable extension [7]. It is the right time and also beneficial to investigate new scalable approaches based on emerging HEVC standards.

Generally, there are three types of scalabilities: temporal, quality, and spatial. HEVC supports the temporal scalability naturally due to the hierarchical prediction structure. The quality scalability can be treated as a special case of spatial scalability with the same resolution in different layers. Hence, in this paper we focus on the HEVC-based SSVC. SSVC com-

Manuscript received April 15, 2012; revised July 20, 2012; accepted August 22, 2012. Date of publication October 5, 2012; date of current version January 8, 2013. This paper was recommended by Associate Editor B. Pesquet-Popescu.

Z. Shi was with Microsoft Research Asia, Beijing 100080, China. He is now with the University of Science and Technology of China, Hefei 230027, China (e-mail: stoneshi@mail.ustc.edu.cn).

X. Sun and F. Wu are with Microsoft Research Asia, Beijing 100080, China (e-mail: xysun@microsoft.com; fengwu@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2012.2223031

presses a video into multiple layers, including one base layer at the lowest resolution and several enhancement layers at incremental higher resolutions. In contrast with the traditional video coding schemes, SSVC introduces a unique problem—how to explore the correlation between layers efficiently so as to achieve comparable coding performance. Usually, there are two ways to exploit the interlayer correlation: pyramid layered approaches and subband based approaches.

Pyramid approaches [8], [9] encode the lower resolution layers first and then utilize the up-sampled texture and motion information coded at lower layers to predict the frames at higher resolution layers. Prior coding standards, such as MPEG-2, H.263, and scalable extension of H.264/AVC, are all such examples. To further improve coding performance, some efforts have been made to enhance the quality of up-sampled reconstruction from lower resolutions for highly efficient interlayer frame prediction. Zhang *et al.* propose to utilize the 2-D Wiener interpolation filter instead of 1-D separate filters to perform better up-sampling [10]. Wu *et al.* propose treating the up-sampling of the video frame as an inverse problem of the initial down-sampling operation and come up with an optimal function to solve this [11]. These schemes improve the quality of the up-sampled signal by performing the inverse filtering according to the total least-square error principle. However, the interlayer prediction in the pyramid approaches is performed from either the corresponding lower layer or the temporal neighboring frames in the same layer. The temporal and spatial correlation need to be exploited in more efficient ways.

The subband approaches [12]–[16] aim to exploit the cross-resolution correlation between neighboring layers by decomposing the input frame into different subbands. Generally, temporal wavelet transformation incorporated with motion compensation is utilized to decompose the temporal correlations between frames. The spatial scalability is usually supported by spatial wavelet transform. Kim *et al.* propose a temporal extension of the set partitioning in a hierarchical tree algorithm (SPIHT) for SVC [13]. Secker *et al.* investigate the 3-D lifting structure for the temporal wavelet transformation [14]. Xiong *et al.* propose a block-size adaptive motion alignment approach in 3-D wavelet coding to further exploit temporal correlation [15]. Li [16] proposes utilizing the over-complete motion prediction to improve the efficiency of motion estimation and compensation in wavelet subbands.

The interlayer correlation can also be exploited in a hybrid way. Similar to subband approaches, interlayer correlation can also be exploited by in-scale motion compensation in the spatial domain [17]. In-scale motion compensation divides a video signal into two different frequency components: low-pass and high-pass components. The low-pass signal uses the lower resolution layers for prediction whereas the high-pass one utilizes the high-pass signals within the same resolution layer for prediction. Though efficient in exploiting the correlations, this scheme still performs at block level and moreover, needs to transmit the motion information, creating additional overhead.

Recently, some efforts have been made to investigate the scalable extension of the emerging video coding standard HEVC. A HEVC-based quality scalable approach is discussed in [18]. The authors propose inserting the base layer reconstruction into the reference list of enhancement layer frames so as to improve the coding performance by increasing the number of references. A more generalized spatial scalability scheme is proposed in [19] in which interlayer prediction mechanisms similar to that of the scalable extension of H.264/AVC are investigated. However, these methods cannot fully take advantage of the techniques adopted by HEVC. The cross-resolution correlation, including both temporal and spatial correlations, should be further exploited in more efficient ways to achieve better coding performance.

In this paper, we propose a HEVC-based SSVC scheme in which two different interlayer prediction modes are proposed: quadtree-based interlayer prediction mode (Q-mode) and learning-based interlayer prediction mode (L-mode). Both SL and multiloop (ML) solutions are supported in our scheme by switching between these two modes.

Q-mode conducts the interlayer prediction based on the quadtree coding structure of HEVC. The quadtree structure plays a key role in HEVC. It contains two kinds of information: prediction information [including coding unit (CU) splits, prediction unit (PU) splits, transform unit (TU) splits, mode, and motion information] and texture information (including residual signals and reconstructed signals). Taking a two-layer coding as an example, Q-mode utilizes the base layer quadtree to predict the enhancement layer quadtree. For this purpose, Q-mode generates a prediction quadtree for each node of the enhancement layer quadtree. The prediction quadtree contains merged information from the corresponding base layer subquadtrees. According to rate-distortion optimization, the enhancement layer quadtree selects the optimal prediction quadtree for each node. The Q-mode only utilizes the prediction information and partial texture information of the base-layer quadtree, which does not demand the full reconstruction of the base layer. Thus, Q-mode supports the SL decoder with limited increase in complexity.

Furthermore, inspired by the learning-based approaches in image hallucination [20], [21], we propose an extra interlayer prediction mode, namely L-mode, to further reduce interlayer redundancy. The L-mode extends the learning-based approach into the SSVC scenario and it exploits temporal and spatial correlations simultaneously through learning-based patch searching at pixel level. A more accurate reconstruction is achieved by mapping all the searched patches together and then used as an interlayer frame prediction within an ML framework. The learning-based patch searching and mapping dose not demand additional overheads.

The rest of this paper is organized as follows. The background on SVC and HEVC is briefly reviewed in Section II. The framework of our proposed SSVC scheme is introduced in Section III. The Q-mode for SL solution is described in Section IV. Section V discusses the extra L-mode in greater detail. The performance of our proposed scheme is evaluated in Section VI. Finally, Section VII concludes the paper.

## II. Background

In this section, we first briefly review the background information on the latest SVC standard—the scalable extension

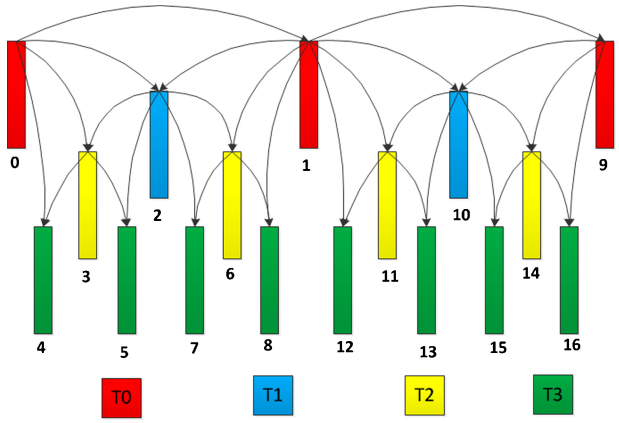



Fig. 2. Hierarchical prediction structure supporting temporal scalability.

of H.264/AVC, and then introduce some basic concepts of the emerging HEVC standard.

### A. Review of the Scalable Extension of H.264/AVC

The scalable extension of H.264/AVC (H.264/SVC) supports three different kinds of scalabilities: temporal scalability, quality or SNR scalability, and spatial scalability.

The temporal scalability is supported by the so-called hierarchical prediction structure as shown in Fig. 2. The continuous pictures are divided into different groups of pictures (GOPs). Pictures in the same GOP are further labeled by different temporal identifiers, e.g., T0 and T1. According to the temporal identifier, the pictures can be divided into different temporal levels. The pictures at the lower levels have higher priority and are usually encoded first. In prediction, a picture can only utilize the pictures in the same or lower temporal levels as its reference pictures. For example, the picture "1" labeled as "T0" can only use the picture "0" as its reference while the picture "2" labeled as "T1" can use both "0" and "1" as its reference as shown in Fig. 2

The quality and spatial scalability can be realized with a multilayer coding structure, specifically the pyramid structure. The quality scalability can be regarded as a special case of spatial scalability with the same picture size for both base and enhancement layers. As shown in Fig. 3, the input sequence is first down-sampled to different resolutions, resulting in different spatial layers. The lowest spatial layer employs a standard H.264/AVC encoder to generate the base layer bitstream. Higher spatial enhancement layers adopt additional interlayer prediction mechanisms to exploit the redundancy between two successive layers to improve coding efficiency.

There are three interlayer predictions in H.264/SVC: the interlayer motion prediction, the interlayer intra prediction, and the interlayer residual prediction. The interlayer motion prediction utilizes the base layer motion information to reduce the motion redundancy between two spatial layers. The interlayer intra prediction can be regarded as the supplement for interlayer motion prediction. When base layer blocks are coded in intra mode and have no motion information, the upsampled reconstruction is directly used for prediction. Moreover, the residual prediction is utilized to further reduce residual redundancy between inter-coded blocks of two successive layers.

These three interlayer predictions cooperating with the traditional inter and intra modes enable a so-called single-loop motion compensation framework. In this framework, all intra-coded blocks in the base layer are coded following the constrained intra prediction [22], so that these intra-coded blocks can be reconstructed without reconstructing any inter-coded blocks. Thus, the decoder does not perform the motion compensation at the base layer, which limits the increase of the decoding complexity. However, the SL solution creates a certain amount of loss of coding efficiency compared with the ML framework as reported in [5].

The H.264/SVC scheme also supports some other coding tools (e.g., MGS coding option) to improve coding performance or to make scalable bitsteam more compatible (e.g., SVC to AVC rewrite mode) to certain scenarios. One can refer to the overview paper [5] for the details.

### B. Basic Concepts of HEVC

The emerging HEVC standard adopts a similar but more complicated coding framework compared with H.264/AVC. New technologies and coding tools are employed to make a significant coding performance improvement. In this section, we briefly introduce some basic concepts of HEVC.

One of the main characteristics of HEVC is the large-size and partition-adaptive coding blocks. Similar to the concept of macroblock in H.264/AVC, HEVC employs large-size blocks, namely treeblocks. The size of a treeblock can be predefined in an sequence parameter set (SPS) and it is usually set to 64. A treeblock can be further split into smaller blocks recursively, which leads to a quadtree based partition. As shown in Fig. 4, $N0$ is a treeblock. It splits incrementally into $N10$ to $N33$ and generates a quadtree. In other words, a quadtree is associated with a treeblock as the root and it splits until a leaf is reached. Each leaf (e.g., $N13$, $N20$, $N32$ in Fig. 4) of the quadtree is referred to as a coding node, called a CU. Each CU can be further split into PUs and TUs as shown in the red circle of Fig. 4. Then the prediction and residue signals are generated based on the partitions of PU and TU, respectively. In HEVC, the whole partitions from treeblock to TU are adaptively determined in a rate-distortion optimal way. Thus, the quadtree-based coding structure helps to achieve more flexible and accurate prediction signals and more compact residuals, especially for high-resolution video sequences.

HEVC also adopts much more complicated ways to exploit the spatial and temporal correlations. The intra prediction adopts up to 34 directions. The inter prediction uses more motion predictors called advanced motion vector predictors (AMVPs). Generally, there are five AMVPs: four spatial MVPs from neighboring CUs (i.e., one Up, one Left, and two Corners) and one additional temporal MVP (TMVP) from collocated CU in the reference frame as shown in Fig. 5.

In addition, HEVC integrates some other new technologies, such as adaptive loop filter (ALF), nonsquare transform, coding with increased bit-depth and so on. Detailed introduction on these new coding tools can be found in [6].

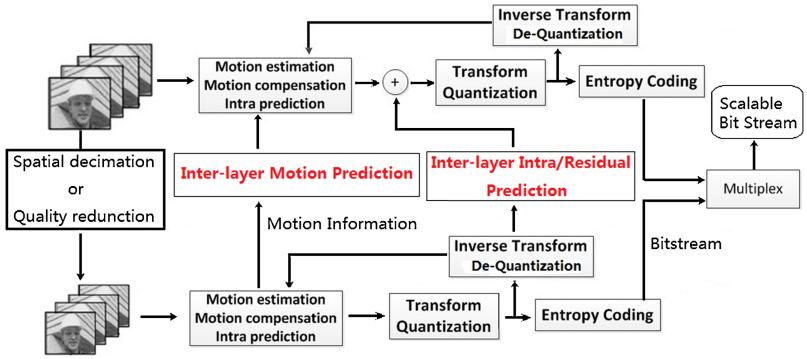



Fig. 3. Framework of the H.264/SVC encoding.

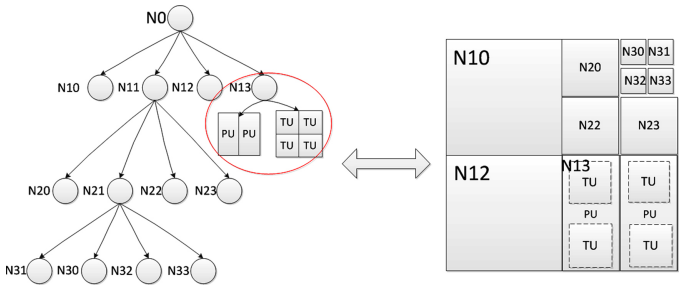



Fig. 4. Quadtree-based coding structure in HEVC. Left: quadtree presentation. Right: plane presentation.

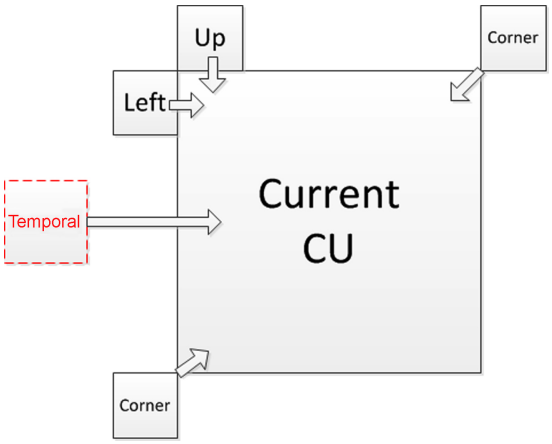



Fig. 5. Adaptive motion vector predictors in HEVC.

## III. Framework of Our Proposed SSVC Scheme

Our proposed SSVC scheme employs the pyramid-based coding structure. It exploits the interlayer correlation by two interlayer prediction mechanisms. In this section, we will introduce the framework of our scheme. The details on the two interlayer prediction methods are discussed in Section IV and Section V, respectively.

For simplicity, we use two-layer encoding structure to explain our proposed SSVC scheme. The interlayer prediction approaches can be easily extended to support the multilayer structure. The framework of our proposed scheme is shown in Fig. 6. There are two spatial layers coded at different resolutions. The input high-resolution sequence is first down-sampled to get the low-resolution sequence. The low-resolution sequence is then coded as the base layer to produce the base layer bitstream. We adopt the standard HEVC encoder to generate the base layer bitstream.

We propose two kinds of interlayer prediction mechanisms to reduce interlayer redundancy. As shown in Fig. 6, they are the basic Q-mode and the extra L-mode. The Q-mode predicts the enhancement layer quadtree from the base layer quadtree. For each node of enhancement-layer quadtree, Q-mode generates an interlayer prediction quadtree containing the coded prediction information (e.g., CU split, PU partition, and motion vectors) and residual information from the base layer subquadtree. Rate-distortion optimization is used to select the optimal interlayer prediction quadtree. The Q-mode can be conducted without full reconstruction of the base layer. Thus it supports the SL decoding by avoiding the motion compensation, deblocking filtering, and ALF in the base layer. This ensures limited increase of complexity compared with the single-layer decoding.

In addition, we propose an extra L-mode to further reduce the interlayer redundancy. It utilizes the learning-based approach to exploit the temporal-spatial correlation simultaneously. In L-mode, we perform learning-based patch searching and mapping among the reference frames from both base layer and enhancement layer. After patch mapping, two refined prediction pictures $F_p^1$ and $F_p^2$ are achieved for further interlayer prediction. Since the patch searching and mapping require the full reconstruction of the base layer, the ML decoder is necessary when enabling L-mode.

Our scheme supports both an SL solution and an ML solution by adaptively using the two interlayer prediction modes. One can easily switch between these two modes to balance the coding performance and complexity according to application requirements.

## IV. Quadtree-Based Prediction Mode

Our proposed Q-mode is introduced in this section. In HEVC, the quadtree coding structure plays a key role in achieving high coding efficiency. It enables adaptive CU splits as well as PU/TU partitions and leads to enhanced predictions and reduced residuals. This adaptive structure, though determined by rate-distortion optimization, is closely connected to the local content of video frames. We observe that the regions

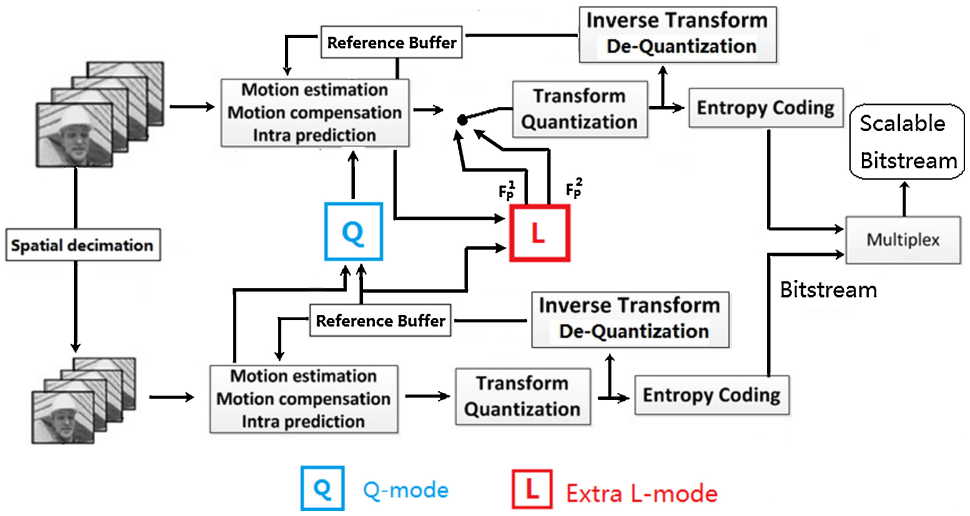



Fig. 6. Framework of our SSVC scheme.

with large motions or complicated textures tend to be coded in smaller blocks and vice versa. Since the enhancement layer frames have corresponding and synchronic contents to the base layer frames, it is reasonable to assume that the base layer quadtrees are able to provide useful structure information for the enhancement layer coding. Accordingly, our Q-mode generates an interlayer prediction quadtree to predict the enhancement-layer quadtree. We further propose an adaptive mechanism to determine the optimal prediction quadtree for each enhancement layer node.

### A. Interlayer Prediction Quadtree

For simplicity, we explain our approach with a spatial ratio of 1:2. The approach can be easily extended to other spatial ratios and is also compatible with quality scalability with a 1:1 ratio. When coding the enhancement layer, the corresponding base layer quadtrees are available. According to the introduction in Section II, it contains three kinds of units: CUs, PUs, and TUs as shown in Fig. 7. The CU splits and PU partitions are related to the prediction signals while the TU partitions are determined by residual signals. In Q-mode, the interlayer prediction quadtree is derived from the corresponding base layer quadtree.

Suppose the root node $N0'$ of enhancement layer quadtree corresponds to a base-layer node $N11$ (size of $32 \times 32$) as shown in Fig. 7. The base layer node $N11$ and its child nodes present a subquadtree from which the interlayer prediction quadtree of $N0'$ is generated. Since the spatial radio is 1:2, the base layer child nodes should be up-sampled first. For example, $N20$ (size of $16 \times 16$) is up-sampled to $N10'$ (size of $32 \times 32$). The PU and TU size is also up-sampled in the similar way. The motion vectors (MVs) included in each PU is up-sampled by multiplying the factor of 2. Once the prediction quadtree is achieved, the prediction signal *Pred* are produced as

$$Pred = \cup_{p_i \in P} MC(MV_{p_i}, Ref_{p_i}) \tag{1}$$

where $P$ is the set of PU partitions, $p_i$ denotes the $i$th partition in $P$, $MV_{p_i}$ and $Ref_{p_i}$ are the motion vectors and reference frame index in $p_i$, $MC(.)$ denotes the motion compensation.

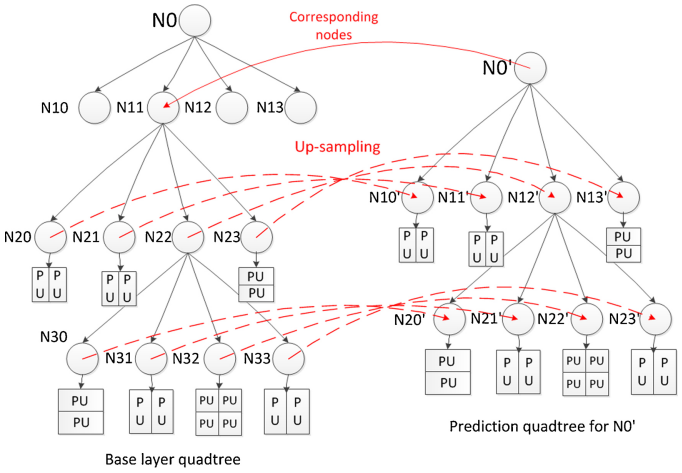



Fig. 7. Quadtree-based interlayer prediction.

According to (1), Q-mode generates the prediction signals for each up-sampled PU in the prediction quadtree. This ensures that the coded base layer motion information is used as much as possible.

There are three kinds of special cases should be addressed.

Case 1: An enhancement layer node may have no corresponding base layer node. For example, $N0$' will have no corresponding base layer node if $N0$ in Fig. 7 does not further split into smaller ones (e.g., the connections between $N0$ to $N10$, $N11$, $N12$, and $N13$ are deleted). In this case, we propose to derive the PU and TU information from the nearest available parent node $N0$. Thus the PU partition of $N0'$ is set equal to the size of $N0'$ and the motion information of $N0'$ derived from the corresponding motion information of $N0'$.

Case 2: The corresponding base layer node is a leaf node. For example, the enhancement layer node $N10'$ corresponds to the base layer leaf node $N20$. In this case, the prediction quadtree is with only one single node. The PU partition and motion information in $N10'$ is achieved by directly up-sampling the corresponding information in $N20$.

Case 3: Some nodes in prediction quadtree are coded in intra mode and the motion information in (1) is not available.

In this case, we employ constrain intra prediction (CIP) [23] in the base layer coding. Due to CIP, the reconstruction of these nodes can be achieved without reconstructing any inter-coded nodes. Then the interlayer prediction signals are achieved by up-sampling the corresponding reconstructions in these intra-coded nodes.

The interlayer prediction quadtree could also contain the base layer TU information. The residual signals from base layer TUs can be up-sampled in spatial domain and then utilized to reduce the interlayer residual redundancy. If the base layer residual signals are used, (1) should be modified as

$$Pred = \cup_{p_i \in P} MC(MV_{p_i}, Ref_{p_i}) + \cup_{t_i \in T} Res_{t_i} \quad (2)$$

where $T$ is the set of TU partitions, $t_i$ denotes the $i^{\text{th}}$ TU in $T$, and $Res_{t_i}$ represents the up-sampled residual signals.

We would like to point out that our prediction quadtree does not include the information on TU partitions. In our scheme, each enhancement layer node selects the optimal TU partition in a rate-distortion optimal way (similar to HEVC). Since our Q-mode employs two kinds of predictions—motion prediction and residual prediction, we set two separate flags in each CU for indicating these two predictions, respectively.

Furthermore, we employ a new TMVP for coding the enhancement layer blocks. The new TMVP is derived from the corresponding up-sampled PU in the prediction quadtree. It is then utilized by conventional inter or merge modes in HEVC.

### B. Adaptive Utilization of Prediction Quadtree

When a node is predicted by the prediction quadtree, all the information including CU split, PU partition, and motion information of this node is derived from that of the base layer. Hence, the prediction quadtree helps to save the bits for coding the split, partition, and motion information. However, the prediction quadtree is not always optimal, especially when very accurate prediction signals are desirable for higher bit rate coding. Given this, we propose adopting the interlayer prediction quadtree adaptively based on the rate-distortion optimization. The optimal interlayer prediction quadtree is selected by minimizing the cost function

$$Cost_{\text{node}} = \min(RD(Q_{\text{mode}}), RD(M), Cost_{\text{Childnodes}}), \quad M \in \{Inter, Intra, Skip, Merge, etc.\} \quad (3)$$

where $Cost_{\text{node}}$ is the minimalized rate-distortion cost of the current node, $RD(Q_{\text{mode}})$ is the rate-distortion cost of the utilization of Q-mode, $RD(M)$ is the rate-distortion cost of the conventional HEVC modes M (M includes the inter, intra, skip, merge, and other supported prediction modes in HEVC), $Cost_{\text{Childnodes}}$ denotes the total rate-distortion cost for encoding its four child nodes.

We calculate (3) recursively from the root node until a child node cannot be further split. In this way, an enhancement layer quadtree selects its optimal prediction way for each node between the interlayer prediction tree and conventional HEVC modes. Accordingly, the interlayer prediction quadtree can be utilized as either a full quadtree if the rood node selects it or several subquadtrees for different child nodes.

## V. Learning-Based Prediction Mode

Generally speaking, there are two ways to generate interlayer prediction signals. One is similar to the Q-mode, where the prediction signals are achieved from the motion compensated reference pictures in the current spatial layer. Its efficiency depends heavily on the accuracy of the up-sampled split information and MVs used in the motion compensation. The other way utilizes the up-sampled base layer reconstruction for prediction due to the synchronous content between two successive layers. However, the up-sampled reconstruction usually lacks important high-frequency information because of down-sampling and compression in the base layer. How to restore the lost high-frequency information is the main problem in generating an efficient interlayer prediction.

Some efforts have been made to tackle this issue. The least-square error based methods are proposed in [10] and [11], treating this problem as an inverse filtering of down sampling. This kind of scheme achieves enhanced reconstruction by performing the inverse filtering at the frame level. However, it only utilizes spatial correlation and frame-level filtering pays little attention to the local characters. In-scale motion compensation is proposed in [17] to exploit the spatial-temporal correlation simultaneously. It separates the interlayer prediction signals into low-frequency parts and high-frequency parts and derives the prediction signals from the up-sampled base-layer reconstructions and by in-scale motion compensation, respectively. However, the in-scale motion compensation is still a block-level scheme and overhead bits for motion information are still needed.

Differently, we propose a patch-based approach to generate the interlayer prediction. Our approach is inspired by the learning-based approach investigated for image hallucination in [20] and [21]. We extend the learning-based approach into the SSVC scenario and propose the L-mode to exploit the spatial-temporal correlation simultaneously at patch level. Our L-mode first partitions the interlayer prediction signals into low-frequency parts and high-frequency parts and then utilizes the relationship between two kinds of patches (high-pass filtered patch and difference patch) to restore the lost high-frequency information. The patch-based searching and mapping exploits the spatial-temporal correlation at pixel level and achieves two refined pictures $F_p^1$ and $F_p^2$ based on the base-layer reconstruction. Finally, the two refined pictures are adaptively selected for interlayer prediction using the rate-distortion optimization decision at block level.

In the following, the basic concepts of the learning-based approach are briefly introduced. The details on L-mode are then discussed from subsection B to E.

### A. Learning-Based Approach

The basic idea of the learning-based approach is to study the relationship of image features at different resolutions and then utilize the learned correlation to approximate the lost high-frequency information during down-sampling and/or compression. The essence of this approach is the learning-based mapping process based on visual patches. As shown in Fig. 8, the learning process builds a training database

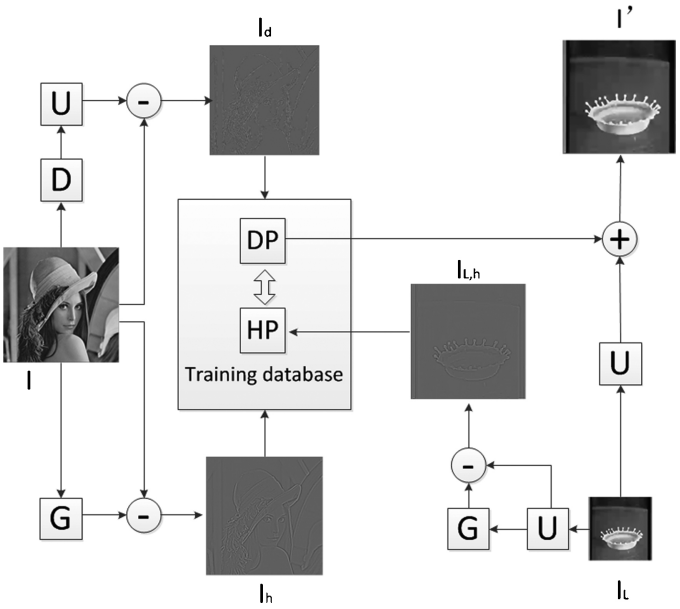



Fig. 8. Learning-based matching.

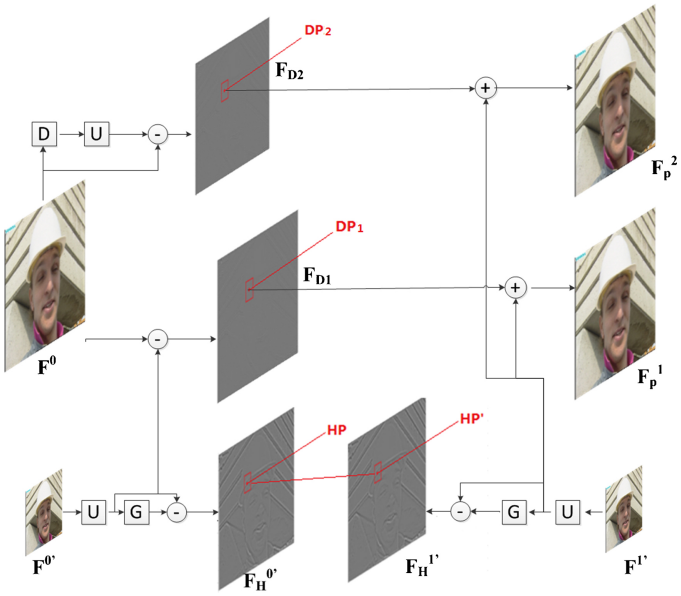



Fig. 9. Visual patch based searching and mapping.

containing visual patch pairs (DP and HP) extracted from an extra image dataset. The mapping process generates high frequency signals by matching between the visual patches in the input image and those in the training database.

Each patch pair in the training database contains two visual patches: the difference patch (DP) and the high-pass filtered patch (HP). The DP is extracted from the difference image $I_d$

$$I_d = I - U(D(I)) \tag{4}$$

where $I$ denotes an original training image, $U(\cdot)$ and $D(\cdot)$ denote the down-sampling and up-sampling, respectively.

One DP has a corresponding HP, which is extracted from the high-pass filtered image $I_h$ at the same location as DP

$$I_h = I - G(I) \tag{5}$$

where $G(\cdot)$ denotes the Gaussian filtering.

According to (4), the DPs contain difference information between the original image and its down-sampled version. In other words, they provide the high-frequency components lost in spatial degradation. According to (5), the HPs are derived from the high-pass filtered image and also contain some high-frequency information. Each HP and DP extracted from the same location forms a pair. As the two paired patches have high visual correlation, the HP can be regarded as a visual hint to the DP. This visual relationship helps us restore the lost high-frequency information. Once DPs and HPs are extracted, they are stored as patch pairs in the training database.

Let $I_{in}$ denote an input image and $I_L$ is the down-sampled version of $I_{in}$. The difference between $I_L$ and $I_{in}$ is estimated in the learning-based approach by maximum-a-posterior as

$$DP' = \arg\max_{DP}(P(DP'|DP, HP \text{ and } HP')) \tag{6}$$

where $DP'$ is the estimated difference patch, $HP'$ is the high-pass filtered patch corresponding to $DP'$ and it is extracted from the high-pass up-sampled image of $I_{L,h}$

$$I_{L,h} = U(I_L) - G(U(I_L)). \tag{7}$$

The learning-based approach solves (6) by searching and mapping processes. The nearest HP is searched from the training database according to the $L_2$ norm between HP and $HP'$. Then, the corresponding DP to the nearest HP is retrieved as $DP'$. It is mapped to the $HP'$ location and integrates into the up-sampled version of $I_L$. The search and mapping process repeats for each pixel until the refined approximation $I'$ is achieved.

### *B. Patch Pairs Construction in L-Mode*

As mentioned previously, the learning-based approach demands prior knowledge of the relationship between the paired patches extracted from offline images. In our SSVC scheme, we do not build an off-line database. Instead, we establish the relationship from the prior coded frames to enhance the correlation used in searching and mapping.

For an interframe $I^1$, there are at least three reference frames available: the corresponding low-resolution reconstruction $F^{1'}$, the high-resolution reference frame $F^0$ and its low-resolution reconstruction $F^{0'}$ as shown in Fig. 9. The high-pass version of $F'$ is derived as

$$F_H^{0'} = U(F^{0'}) - G(U(F^{0'})). \tag{8}$$

And the two different frames are derived as

$$F_{D1} = F^0 - U(F^{0'}) \tag{9}$$

$$F_{D2} = F^0 - U(D(F^0)). \tag{10}$$

As shown in Fig. 9, we build the relationship directly based on $F_H^{0'}$, $F_{D1}$, and $F_{D2}$ rather than building a database to store the extracted patch pairs as in [24]. We notice that building a database requires too much memory buffer. Taking a CIF (352 $\times$ 288) sequence as an example, the buffer size for storing the

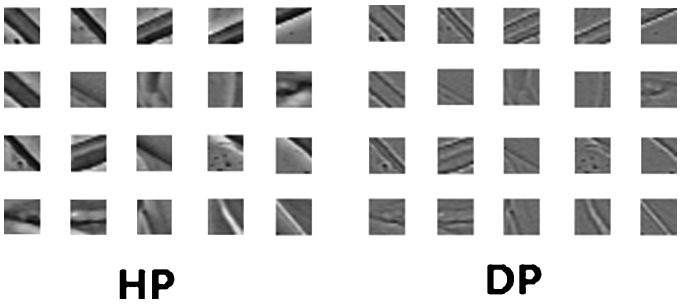



Fig. 10. Visual comparison between patch pairs. Left: high-pass filtered patch. Right: difference patch.

$11 \times 11$ patch pairs is 37 MB. Instead, the buffer requirement for these three frames is limited and only 304 KB.

Our L-mode employs two kinds of patch pairs for prediction. For a given pixel position $(x, y)$, the HP is a $11 \times 11$ square region centering at $(x, y)$ inside $F_H^{0'}$. The corresponding $DP_1$ and $DP_2$ are the square regions centered at the same location $(x, y)$ inside $F_{D1}$ and $F_{D2}$, respectively. Then two kinds of patch pairs, the patch pair (HP, $DP_1$) and the patch pair (HP, $DP_2$), are utilized in our L-mode.

Different from [24], we propose the new patch pair (HP, $DP_2$) for prediction so as these two patch pairs can complement one another in some special cases. A typical case here is the $DP_1$ contains less high frequency than $DP_2$ when the corresponding regions in $F^0$ and the up-sampled $F^{0'}$ are more similar or vice versa. Using two patch pairs increases the possibility of retrieving enough high frequency information for interlayer prediction.

Fig. 10. shows some samples of extracted patch pairs from test sequence *Foreman*. In Fig. 10, the left part shows extracted HPs and the right part denotes the DPs. One can observe that the HPs look quite similar to DPs. Hence, HPs can be regarded as fine approximations to DPs. This similarity between patch pairs is utilized in patch-based searching and mapping.

Moreover, we would like to point out that we only select one reference frame and its low-resolution version to generate the patch pairs although the current frame may have several reference candidates. HEVC adopts the quantization parameter (QP)-fluctuation coding structure hence these candidate reference frames have different QPs. Among all these reference frames, we select the one with smallest QP since it has the best quality and provides the most high-frequency information.

### C. Patch Searching Assisted by Motion Vectors

Due to the highly relevant temporal correlation, we can use HPs extracted from the current up-sampled reconstruction to search the nearest corresponding HPs from the previous reference frames. Based on the visual similarity between patch pairs, we can further estimate the lost high frequency via corresponding DPs of patch pairs {(HP, DP)}.

The patch searching can be performed in the following three steps.

Firstly, the high-pass version of the current base layer reconstruction $F^{1'}$ is generated as

$$F_H^{1'} = U(F^{1'}) - G(U(F^{1'})). \tag{11}$$

Secondly, we utilize the motion information coded in the base layer to constrain the search center for each high-pass patch extracted from $F_H^{1'}$. Let HP′ denote a high-pass patch extracted from $F_H^{1'}$ at position $(x, y)$ and $(MX, MY)$ denote the MV information of location $(x, y)$ which is derived from the up-sampled base layer motion information. The center of the search range of HP is determined by $(MX', MY')$

$$(MX', MY') = \begin{cases} (MX,\ MY), & if\ F_R = F^0 \\ (MX,\ MY) \times \dfrac{|CurrPOC - POC_2|}{|CurrPOC - POC_1|}, & otherwise \end{cases}, \tag{12}$$

where $F_R$ denotes the reference frame signaled by base layer MV information at $(x, y)$, currPOC denotes the picture order count(POC) defined in HEVC indicating the picture index) of current frame $I^1$, $POC_1$ and $POC_2$ are the POCs of $F_R$ and $F^0$, respectively. In other words, the base layer MV need to be scaled if the reference frame $F_R$ signaled by this MV is different from $F^0$. Accordingly, we set the search center $(x',\ y')$ of HP′ as

$$(x', y') = (x + MX',\ y + MY'). \tag{13}$$

Finally, the location (*PosX, PosY*) of the nearest HP to HP′ is achieved subject to

$$(PosX,\ PosY) = \arg \min_{(x,y) \in R} (SAD(HP(x, y), HP')) \tag{14}$$

where $SAD(\cdot)$ denotes the sum of the absolute difference between two patches and R denotes a search region centered at $(x', y')$ with size $n \times n$ in $F_H^{0'}$.

### D. Patch Mapping

Once the location (*PosX, PosY*) is decided through the searching process, the corresponding patch pairs (HP, $DP_1$) and (HP, $DP_2$) also become available. The mapping process utilizes $DP_1$ and $DP_2$ to retrieve the fine prediction frame $F_p^1$ and $F_p^2$.

We perform the mapping process pixel by pixel based on the up-sampled reconstruction of $F^{1'}$. Taking $F_p^1$ as example, for each pixel, a corresponding DP1 is achieved by the searching process. All these retrieved patches are integrated with each other, generating the estimated high frequency signal $P_1$ by

$$P_1(x, y) = \frac{1}{N} \sum_{i=1}^{N} DP_1^i(x, y) \tag{15}$$

where $N$ is the number of overlapped patches at position $(x,\ y)$.

Together with the up-sampled frame $F_U (F_U = U(F^{1'}))$, the final prediction frame $F_p^1$ is achieved by

$$F_p^1(x, y) = F_U(x, y) + P_1(x, y). \tag{16}$$

Similarly, $F_p^2$ is generated by integrating all the retrieved $DP_2$ patches to $F_U$.

In a word, the prediction frames in L-mode consist of two parts: the low-frequency part $F_U(x, y)$ from up-sampled base-layer reconstruction and the high-frequency part $P(x, y)$ derived from patch-based mapping.

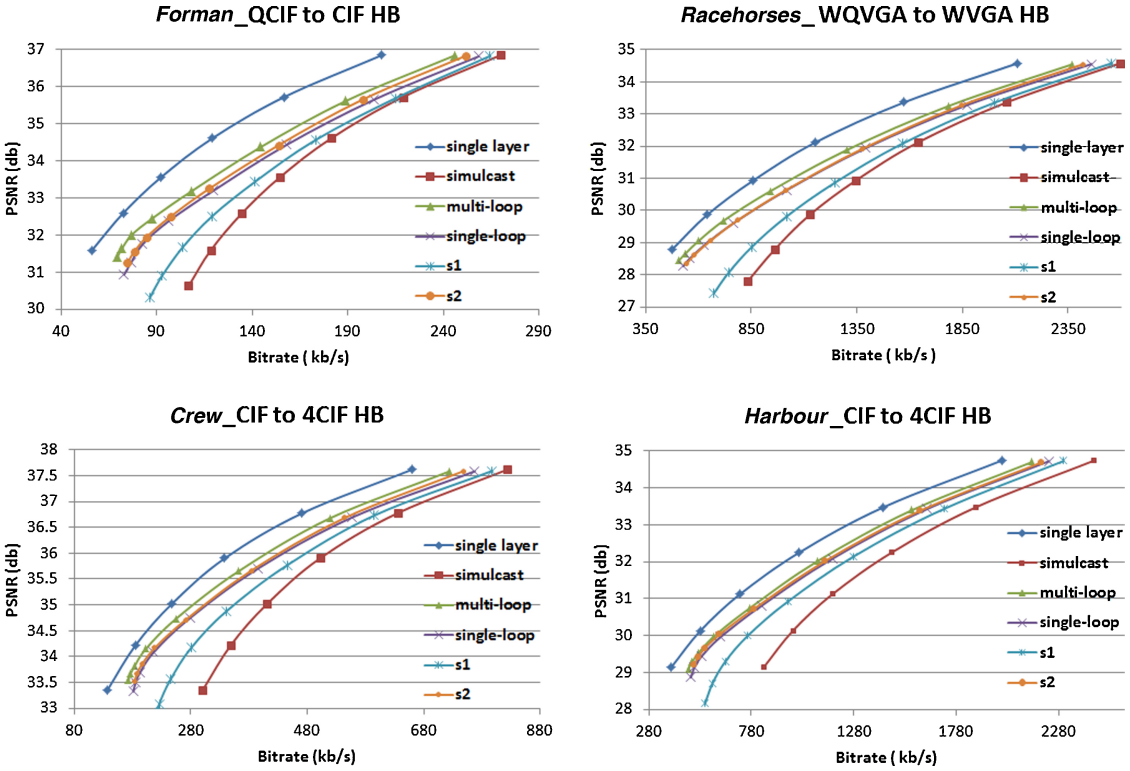



Fig. 11. Performance comparison in HB coding with BL_QP = 30.

### *E. Adaptive Utilization of L-Mode*

Although the learning-based patch mapping is performed at pixel level, the utilization of the refined prediction frames are at block level so that it can be compatible with HEVC quadtree based coding structure. In L-mode, the PU partition is set equal to the size of the current CU. The prediction signals are directly derived from the collocated regions in the two refined prediction frames: $F_p^1$ and $F_p^2$. The TU partition is determined by the HEVC rate-distortion optimal decision.

The rate-distortion optimized selection in (3) is slightly modified into (17) to support the extra L-mode

$$Cost_{\text{node}} = \min\left(RD\left(F_p^1\right), RD\left(F_p^2\right), RD(M), Cost_{\text{Childnodes}}\right), \quad \text{where } M \in \{Q_{\text{mode}}, \textit{Inter}, \textit{Intra}, \textit{Skip}, \textit{Merge}, \textit{etc.}\} \tag{17}$$

Here, $RD\left(F_p^1\right)$ and $RD\left(F_p^2\right)$ are the rate-distortion cost for utilizing $F_p^1$ and $F_p^2$ as prediction signals, respectively.

In this way, the L-mode can be adaptively utilized in a HEVC-compatible SSVC encoder. We would like to point out that L-mode has already utilized the temporal motion information coded in the based layer via the patch-based searching and mapping. Thus we modify the Q-mode slightly in which the motion information is generated from the conventional motion estimation within the same spatial layer instead of from the up-sampled based-layer. In other words, Q-mode can be regarded as a conventional inter mode cooperated with interlayer residual prediction when extra L-mode is enabled.

The advantage of our L-mode can be summarized in three aspects. First, the temporal-spatial correlation is exploited simultaneously. The base-layer reconstruction provides a coarse spatial prediction for enhancement frames and the temporal correlation is utilized through the learning-based mapping. Second, the refinement is performed at pixel level. This helps to enhance the accuracy of the prediction signals compared with frame-based or block-based approaches. Third, the temporal-spatial correlation is exploited according to the visual relationship between patch pairs. All these patches can be achieved on the decoder side. Thus, no additional overhead in introduced into the bitstream.

### *F. Complexity*

This section discusses the decoding complexity of our proposed scheme. Our base-layer coding is compliant to HEVC so that the decoding complexity at the lowest resolution is the same as that of HEVC. The complexity of coding the higher resolution layers is adaptable. The SL decoding is supported if only Q-mode is enabled as in this case only the intra-coded CUs need to be reconstructed while motion compensation for all the inter-coded CUs are skipped in the base-layer decoding. Thus, the complexity increase in the SL solution is limited compared with the single-layer decoding.

When L-mode is enabled, the ML decoding is required. The complexity increase comes from two factors: one is the motion compensation at the base layer and the other is the learning-based patch searching and mapping. The later one involves three modules: up-sampling, high-pass filtering, and patch-searching. The complexity of up-sampling is determined by the up-sampling filters. In our scheme, we adopt 1-D separate interpolation filters as same as those used in JSVM [26]. The length of the filters is $L$. For a frame of size $N \times N$,

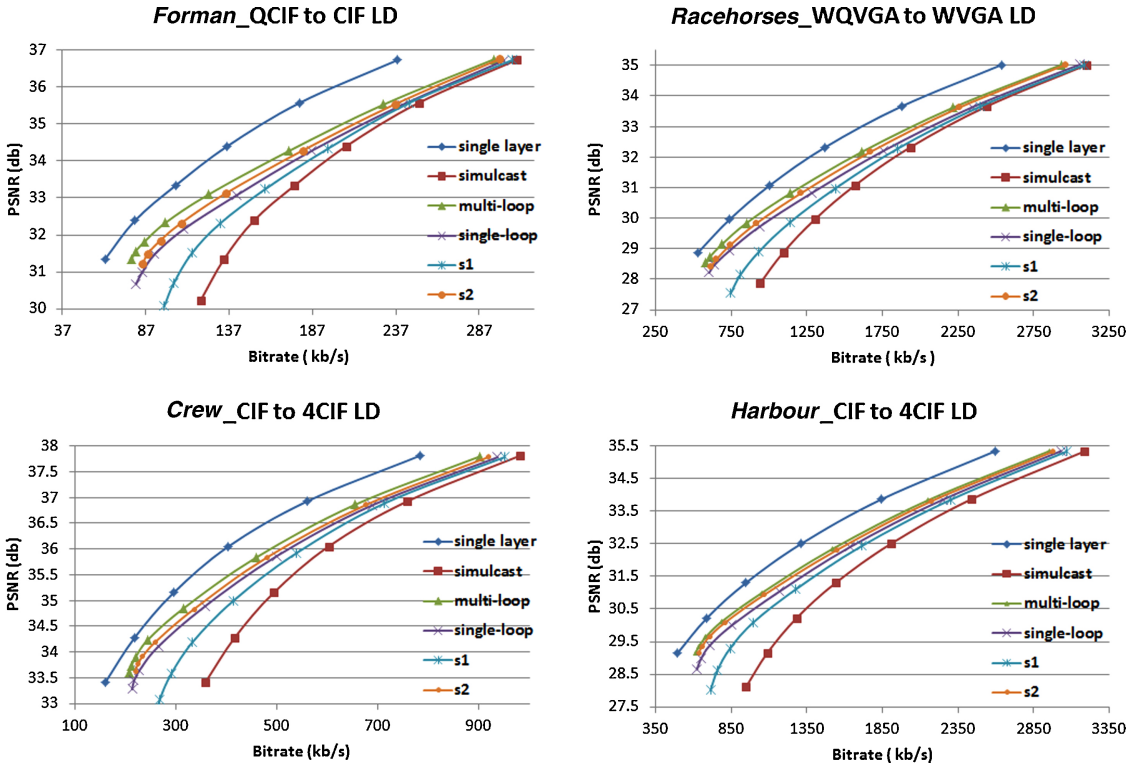



Fig. 12. Performance comparison in LD coding with BL_QP = 30.

the complexity of up-sampling is $O(L * N^2)$. The high-pass filtering needs convolution operation with a Gaussian kernel. If the size of the kernel is $M \times M$, the complexity of high-pass filtering is $O(M^2 * N^2)$. Since the patch searching process is performed at pixel level, its complexity is determined by the patch size $P$, search region $R$, and number of pixels. The total complexity is $O(P^2 * R^2 * N^2)$. However, we would like to point out that the search process in our scheme supports parallel implementation. The pixels can be divided into several groups according to the computational power of a multithread or multicore system. The complexity can be greatly reduced though parallel processing. The complexity of L-mode can be reduced to $O((L + M^2 + P^2 + R^2) * N^2/\beta)$ if the parallel factor is $\beta$.

## VI. Experimental Results

We implement our proposed schemes in HM3.0 [25] and perform extensive experiments to evaluate the performance in comparison with simulcast, single-layer, and the state-of-the-art SSVC schemes [17], [19].

In the experiments, two spatial layers [one base layer (BL) and one enhancement layer (EL)] are encoded with a spatial radio of 1:2. The QPs used in two layers are denoted as BL_QP and EL_QP, respectively. For each video sequence, BL_QP is fixed whereas EL_QP is set to a series of different values so that a PSNR-rate curve of EL is achieved. The input videos for BL coding are generated from EL input using down-sampling tool provided in JSVM [26]. The 1-D separate filters provided by H.264/SVC [5] are adopted for up-sampling in the tests. For the SSVC schemes, the BL is encoded by the HEVC encoder with high efficiency configuration [6] and the EL encoder adopts a HEVC-compatible encoder supporting different interlayer prediction modes. For the simulcast solution, two layers are both coded by HEVC without the interlayer prediction.

For comparison, we also implement two additional HEVC-based SSVC schemes: an SL scheme as proposed in [19] and an ML scheme using in-scale motion compensation as proposed in [17]. These two schemes are labeled as "S1" and "S2" in our experimental results, respectively. Notice that our proposed Q-mode is used in the ML solution S2. The base layer motion vector is also utilized as an additional motion vector predictor for the in-scale motion compensation.

### A. Coding Performance

We report the overall coding performance in this subsection. We evaluate the performance using the two coding structures suggested by HEVC, the hierarchical-B (HB) coding and low-delay (LD) coding, respectively.

Fig. 11 shows the PSNR-rate comparisons with HB coding case for four test sequences. The GOP size is 8. The BL_QP is set to 30 and the EL_QP changes from 28 to 42 with an interval of 2. The curve denoted as "ML" denotes coding performance of the ML solution with the extra L-mode enabled. The curved marked as "SL" denotes the performance of SL solution by only enabling Q-mode. It can be observed that the ML solution achieves 27.48% Y-BD-rate saving in average, where L-mode and Q-mode contribute 25.13% and 2.35% Y-BD-rate saving, respectively. The SL solution has some coding efficiency loss, which achieves 22.73% BD-rate saving by only enabling Q-mode. The loss mainly comes from two factors: first, the CIP brings some performance loss for each layer

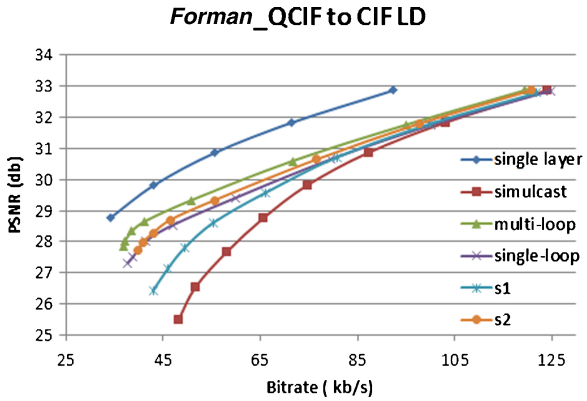



Fig. 13. Performance in LD coding with BL_QP=37.

coding [23]; second, the base layer reconstruction cannot be used for the inter-coded blocks in the enhancement layer which reduces the efficiency of interlayer prediction.

Similarly, the performance of low-delay (LD) coding is evaluated in Fig. 12. The ML solution brings 25.05% Y-BD-rate saving, in which L-mode and Q-mode contribute 22.75% and 2.30% Y-BD-rate saving, respectively. Also, the SL solution achieves 18.83% Y-BD-rate saving. Fig. 13 shows another LD result of *Foreman* by setting the BL_QP to 37 and EL_QP from 35 to 49. Similar improvement is achieved in this case.

We observe that our SL solution outperforms the sing-loop scheme S1 in both HB and LD coding cases. This is because our Q-mode not only utilizes the motion and residual information but also the CU split and PU partition information from the base layer. Instead, S1 does not fully take advantage of the coded quadtree in the base layer. Similarly, our ML scheme outperforms the in-scale motion compensation scheme S2 at both HB and LD cases. This demonstrates that the pixel-level learning-based mapping exploits the temporal and spatial correlations better than the block-based in-scale compensation.

We also evaluate our performance in comparison with the simulcast for three HEVC test classes: Class B (1080P), Class C (WVGA), and Class D (WQVGA). The BD-rate results are shown in Table I. Our ML scheme achieves 3.46% Y-BD-rate saving compared with ML S2 averagely and our SL scheme shows up to 17.38% Y-BD-rate gain compared with SL S1. Besides, our ML solution outperforms SL solution with a 4.92%Y-BD-rate saving on average for these three test classes.

Furthermore, we compare our ML solution with an overlapped block motion compensation (OBMC) based S2 scheme, where the OBMC is integrated into the in-scale motion compensation as suggested in [27]. Based on our observation, the OBMC can bring additional 0.25% and 0.23% Y-BD-rate savings for Class C and Class D sequences in the LD case, respectively. Although OBMC involves an overlapped mechanism similar to our patch mapping, our L-mode outperforms the OBMC based S2 because in-scale motion compensation, with or without OBMC, requires motion information for each block whereas our L-mode achieves the refined prediction frame without any additional overheads.

It should be noted that our coding gain in the LD case is relatively lower than that in HB case for some test sequences.

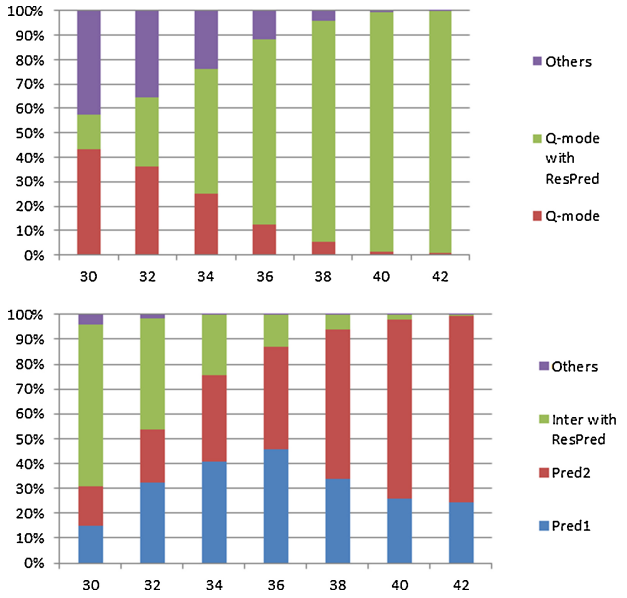



Fig. 14. Prediction mode distribution at BL_QP = 30 (up: only enabling Q-mode, bottom: enabling extra L-mode).

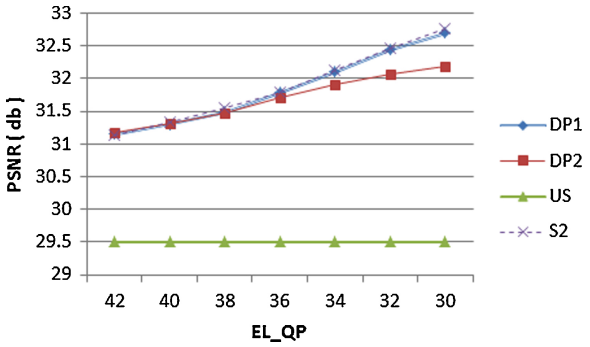



Fig. 15. Comparison of predictions (for *Foreman*).

This is because temporal correlation is stronger in LD. It diminishes the spatial correlation between two layers and thus reduces the efficiency of interlayer prediction.

### B. Distribution of Prediction Modes

In this section, the distribution of prediction modes in EL layer coding in our proposed scheme is investigated and shown in Fig. 14. Here $y$-axis denotes the percentage of each mode and $x$-axis shows the EL_QP values. The BL_QP is set to 30. "Pred1" and "Pred2" denote that prediction signals are derived from $F_p^1$ and $F_p^2$ in L-mode, respectively.

When only enabling Q-mode, we observe that the Q-mode with residual prediction dominates especially at lower bitrates. As the quality of the enhancement layer become higher, the correlation from the interlayer prediction quadtree becomes lower, thus the percentages of conventional inter and intra modes increase. When further enabling the extra L-mode, we observe that the Pred1 and Pred2 dominate at low bit rates because of better prediction quality achieved using the L-mode. Moreover, Pred2 has more percentage than Pred1 when the quality of the enhancement layer decreases. With the bit rate increasing, the inter prediction becomes better

TABLE I
BD-RATE SAVING IN COMPARISON WITH SIMULCAST FOR CLASS B, CLASS C, AND CLASS D

| | HB (BD-rate saving) | | | LD (BD-rate saving) | | |
|---|---|---|---|---|---|---|
| | Class B (Y/U/V) | Class C (Y/U/V) | Class D (Y/U/V) | Class B (Y/U/V) | Class C (Y/U/V) | Class D (Y/U/V) |
| ML | −24.10/−28.56/−25.76 | −31.20/−29.65/−29.95 | −23.63/−24.6/−20.35 | −23.88/−23.54/−20.04 | −31.23/−28.93/−38.50 | −22.40/−33.00/−25.85 |
| SL | −21.74/−39.26/−24.22 | −26.58/−26.28/−26.28 | −19.15/−18.48/−19.45 | −18.78/−22.34/−19.12 | −23.55/−24.63/−34.45 | −17.10/−21.90/−20.93 |
| S1 | −8.40/−6.90/−8.58 | −9.20/−8.93/−9.13 | −8.40/−7.05/−8.55 | −9.00/−9.02/−10.08 | −10.53/−9.33/−8.68 | −8.75/−11.10/−10.08 |
| S2 | −22.04/−26.86/−25.02 | −27.70/−26.40/−28.65 | −19.60/−21.03/−16.20 | −20.14/−31.60/−20.18 | −27.45/−24.83/−33.18 | −18.78/−28.63/−26.50 |

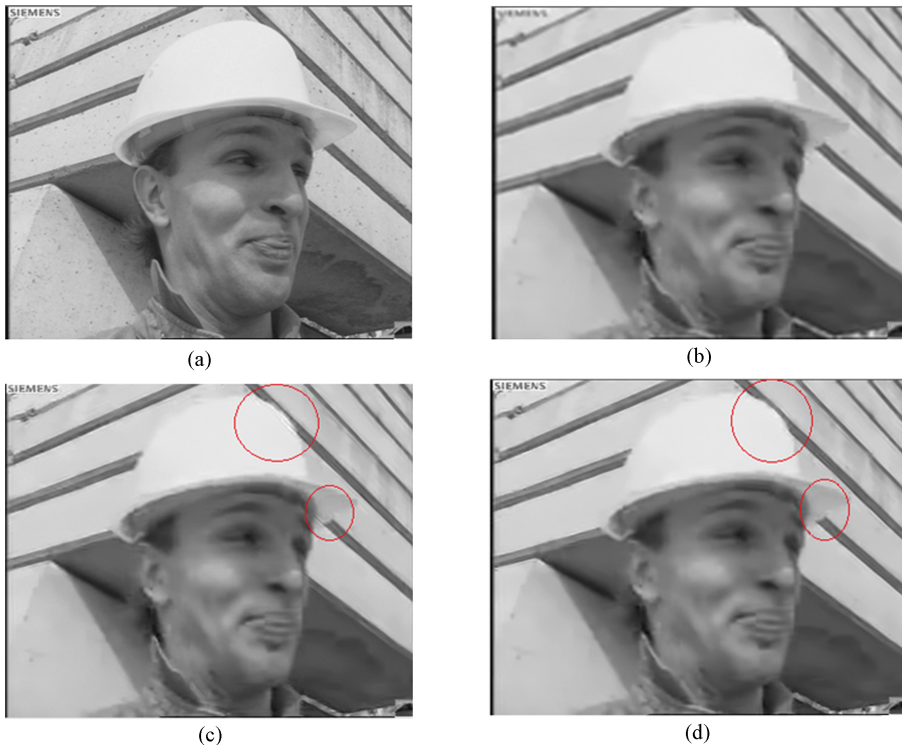



Fig. 16. Visual comparison of prediction frames. (a) Original image. (b) Directly up-sampled (29.64 dB). (c) In-scale motion compensation (32.03 dB). (d) $F_p^1$ in our L-mode (32.16 dB).

than the refined prediction from BL, leading to an increasing percentage of inter mode at higher bit rates.

### C. Improvement in Prediction

In this section, we demonstrate the effectiveness of learning-based refinement for interlayer prediction in Fig. 15. The proposed learning-based approach is performed globally for the whole base-layer reconstruction at each pixel. The labels 'DP1' and 'DP2' denote the prediction frames $F_p^1$ and $F_p^2$, respectively. Another two prediction approaches are implemented for comparison. Here label 'US' denotes the prediction generated by directly up-sample the base-layer reconstruction and label 'S2' corresponds to the prediction produced using the in-scale motion compensation scheme with OBMC.

In Fig. 15, one can observe that the quality of "US" prediction is fixed since the BL_QP value is fixed. With the decrease of EL_QP, the quality of the predictions of the other three methods increases. Among the three prediction methods (DP1, DP2, and S2), DP2 losses at high bit rates but provides slightly better prediction at low bit rates; DP1 and S2 achieves better predictions at high bit rates which outperform US method more than 2 dB; S2 is slightly better than DP1 at high bit rates but requires the overhead bits on motion information.

Fig. 16 demonstrates the visual comparison of these three prediction approaches. The BL_QP and EL_QP are 30 and 36, respectively. According to the results, up-sampling the base layer directly with spatial interpolation filters shows the worst quality. Some high-frequency components can be restored from temporal reference frames by in-scale motion prediction. Our scheme shows better visual quality at some edge regions as denoted by the red circles.

Here, we would like to point out that the in-scale motion prediction, with or without OBMC, requires encoding the corresponding motion information into bitstream whereas our learning-based refinement requires only an additional flag. Even though the prediction quality of the in-scale motion prediction seems similar or slightly higher than that of our scheme, the overall performance of our scheme is higher by

providing a comparable prediction and meanwhile introducing much less overhead bits.

## VII. Conclusion

This paper presents a SSVC scheme based on the emerging HEVC. The proposed SSVC scheme supports both SL and ML solutions by enabling different interlayer prediction mechanisms. For the SL solution, we propose the quadtree-based Q-mode to reduce the interlayer redundancy in terms of the quadtree information. We further propose the learning-based L-mode to improve the coding performance within an ML framework. The L-mode exploits the temporal-spatial correlation simultaneously by patch-based searching and mapping. The high frequency details lost in the base-layer reconstruction are estimated at pixel level, leading to two enhanced references for interlayer prediction. These two interlayer prediction modes are adaptively utilized in our SSVC scheme according to the quadtree-based rate-distortion selection in HEVC. Experimental results demonstrate the effectiveness of our proposed SSVC scheme compared with the other HEVC based SSVC schemes.

To further improve the coding performance of our SSVC scheme, we would like to put efforts on the following works in the future. First, advanced learning and mapping methods should be investigated to enhance the quality of the refined prediction frames especially at high bit rates. Second, the quadtree structure and the correlation between two layers should be further exploited to reduce the overhead bits of enhancement layer. Third, joint optimization in terms of determining both motion vectors as well as quad-tree structures should be further studied with regard to both the base layer and enhancement layer coding.

## References


[1] *Generic Coding of Moving Pictures and Associated Audio Information: Video*, ISO/IEC 13818-2:1996 (MPEG-2 Video), ISO/IEC ITU-T, Jul. 1995.

[2] *Video Coding for Low Bit Rate communication*, ITU-T Rec. H.263, ITU-T, Version 1: Nov. 1995, Version 2: Jan. 1998, Version 3: Nov. 2000.

[3] *Coding of Audio-Visual Objects—Part 2: Visual*, ISO/IEC 14492-2(MPEG-4 Visual), ISO/IEC JTC 1: Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.

[4] *Advanced Video Coding for Generic Audiovisual Services*, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC JTC 1, Version 8 (including SVC extension): consented in Jul. 2007.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable coding extension of H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[6] B. Bross, W. J. Han, J. R. Ohm, G. J. Sullivan, and T. Wiegand, *WD6: Working Draft 6 High Efficiency Video Coding*, JCTVC-H1003, JCT-VC Meeting, Feb. 2012.

[7] *Joint Preliminary Call for Proposals on Scalable Video Coding Extensions of High Efficiency Video Coding (HEVC)*, Geneva, Switzerland, May 2012.

[8] M. Flierl and P. Vandergheynst, "An improved pyramid for spatially scalable video coding," in *Proc. ICIP*, vol. 2. 2005, pp. 878–881.

[9] T. Wang, C. S. Park, J. H. Kim, M. S. Yoon, and S. J. Ko, "Improved interlayer intra prediction for scalable video coding," in *Proc. TENCON*, 2007, pp. 1–4.

[10] W. Zhang, A. Men, and P. Chen, "Adaptive interlayer intra prediction in scalable video coding," in *Proc. ICIP*, 2009, pp. 876–879.

[11] X. Wu, M. Shao, and X. Zhang, "Improvement on H.264 SVC by model-based adaptive resolution upconversion," in *Proc. ICIP*, 2010, pp. 4205–4208.

[12] N. Adami, A. Signoroni, and R. Leonardi, "State-of-the-art and trends in scalable video compression with wavelet based approaches," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1238–1255, Sep. 2007.

[13] B.-J. Kim, Z. Xiong, and W. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical tree (3-D SPIHT)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 8, pp. 1374–1387, Dec. 2003.

[14] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Trans. Image Precess.*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.

[15] R. Xiong, F. Wu, S. Li, Z. Xiong, and Y.-Q. Zhang, "Exploiting temporal correlation with block-size adaptive motion alignment for 3-D wavelet coding," in *Proc. SPIE Visual Commun. Image Process.*, vol. 5308. 2004, pp. 144–155.

[16] X. Li, "Scalable video compression via over-complete motion compensated wavelet coding," *Signal Process.: Image Commun.*, vol. 19, no. 7, pp. 637–651, 2004.

[17] R. Xiong, J. Xu, and F. Wu, "In-scale motion compensation for spatially scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 2, pp. 145–158, Feb. 2008.

[18] Z. Shi, X. Sun, and J. Xu, "CGS Quality Scalability for HEVC," in *Proc. MMSP*, 2012, pp. 1–6.

[19] H. Choi, J. Nam, D. Sim, and I. V. Bajic, "Scalable video coding based on high efficiency video coding (HEVC)," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 2011, pp. 346–351.

[20] J. Sun, N. N. Zheng, H. Tao, and H. Y. Shum, "Image hallucination with primal sketch priors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, vol. 2. 2003, pp. 729–736.

[21] Z. Xiong, X. Sun, and F. Wu, "Image hallucination with feature enhancement," in *Proc. IEEE Comput. Soc. Conf. Computer Vision Pattern Recog.*, 2009, pp. 2074–2081.

[22] H. Schwarz, D. Marpe, and T. Wiegand, "Constrained inter-layer prediction for single-loop decoding in spatial scalability," in *Proc. ICIP*, vol. 2. 2005, pp. 870–873.

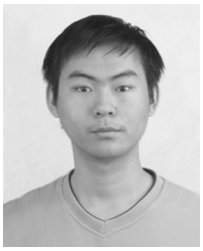
[23] V. Wahadaniah, C. S. Lim, and S. M. T. Naing, *Constrained Intra Prediction Scheme for Flexible-Sized Prediction Units in HEVC*, JCTVC-D094, Daegu, Korea, Jan. 2011.

[24] X. Sun and F. Wu, "Classified patch learning for spatially scalable video coding," in *Proc. ICIP*, 2009, pp. 2301–2304.

[25] *HM3.0* [Online]. Available: http://hevc.kw.bbc.co.uk/trac/

[26] *JSVM 9* [Online]. Available: http://ip.hhi.de/imagecom_G1/savce/downloads/

[27] C.-C. Chen, Y.-Y. Chen, C.-L. Lee, W.-H. Peng, and H.-M. Hang, *CE2: Report of OBMC with Motion Merging*, JCTVC-F049, Turin, Italy, Jul. 2011.

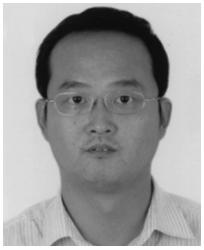
**Zhongbo Shi** received the B.S. degree in electronic engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2009, where he is currently pursuing the Ph.D. degree in electronic engineering.

He has been a Research Intern with Microsoft Research Asia, Beijing, China, since 2010, where his research has been focused on image and video compression, image representation, and image restoration. His current research interests include vision-based and cloud-based image and video compression.

**Xiaoyan Sun** (M'04–SM'10) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1997, 1999, and 2003, respectively.

Since 2004, she has been with Microsoft Research Asia, Beijing, China, where she is currently a Lead Researcher with the Internet Media Group. She has authored or co-authored more than 50 journal and conference papers, ten proposals to standards. She has filed seven granted patents. Her current research interests include vision-based and cloud-based image and video compression

Dr. Sun was a recipient of the Best Paper Award of the IEEE Transactions on Circuits and Systems for Video Technology in 2009.



**Feng Wu** (M'99–SM'06–F'12) received the B.S. degree in electrical engineering from the University of Xi'an Electrical Science and Technology, Xi'an, China, in 1992, and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1996 and 1999, respectively.

He joined Microsoft Research Asia, Beijing, China, as an Associate Researcher in 1999. He has been a Researcher with Microsoft Research Asia since 2001 and is currently a Senior Researcher/Research Manager. He has authored or co-authored more than 200 high-quality papers and filed 67 U.S. patents. His 13 techniques have been adopted into international video coding standards. His current research interests include image and video compression, media communication, and media analysis and synthesis.

Dr. Wu received the Best Paper Award of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2009, PCM 2008, and SPIE VCIP 2007. He was elected an IEEE fellow for his contributions in visual data compression and communication.