Low-Overhead Content-Adaptive Spatial Scalability for Scalable Video Coding

Chia-Wen Lin, *Senior Member, IEEE*, Chia-Ming Tsai, and Po-Chun Chen

***Abstract*—To support spatial scalability, the scalable extension of H.264/AVC (SVC) uses video cropping or uniform scaling to downscale the original higher-resolution (HR) sequence to a lower resolution (LR) sequence. Both operations, however, will cause critical visual information loss in the resized frames. To address the problem, we propose a low-overhead content-adaptive spatial scalability SVC (CASS-SVC) coder consisting of three main modules: a mosaic-guided video retargeter, a side-information coder, and a non-homogeneous inter-layer predictive coder. The proposed video retargeting scheme first constructs a panoramic mosaic for each video shot to obtain a compact shot-level global scaling map which is then used to derive the scaling maps of individual frames in the shot. The side information required for the non-homogeneous scaling, including the global scaling maps and the spatial corresponding positions of individual frames to the panoramic mosaic, are then efficiently coded by the side-information coder. The non-homogeneous interlayer prediction coding tools are used to provide good predictions to reduce the bitrates for coding the HR frames. Our simulation results demonstrate that, compared to existing CASS-SVC coders, our method cannot only well preserve subjective quality of important content in the LR sequence, but also significantly improves the coding efficiency of HR sequence.**

***Index Terms*—Inter-layer prediction, scalable video coding, spatial scalability, video adaptation, video retargeting.**

I. Introduction

NETWORK environments usually involve heterogeneous devices with various display abilities and channel bandwidths. While streaming a video through networks, video content needs to be adapted to match the heterogeneity of networks and user devices. Salable video coding, e.g., the scalable extension of H.264/AVC (SVC) [1], is an important technology to support the video content adaptation. More specifically, to accommodate the different resolutions and aspect ratios for different types of display devices such as standard-definition TVs (SDTVs), high-definition TVs (HDTVs), home theater projectors, and hand-held devices, the spatial scalability offered in SVC [1], [2] is a useful tool.

In order to support various display resolutions and aspect ratios, SVC supports video cropping or uniform scaling to downscale the original higher-resolution (HR) sequence to a lower resolution (LR) sequence. Different resolution videos are coded by individual video encoders and interlayer prediction are then used to reduce the redundancies between different spatial layers [2]. However, the flexibility and performance of the spatial scalability in SVC is still rather limited, since both video cropping and uniform scaling used in SVC lead to critical visual information loss. Recently, several content-adaptive video retargeting methods have been proposed [4]–[11]. These methods mainly aim to retain as much human interested regions as possible by trimming unimportant spatio-temporal content, thereby preserving in the resized video the main content inside the source video. Therefore, they can be used to help enhance the flexibility and performance of current SVC.

According to the definition in [3], content-adaptive video retargeting methods can be classified into discrete approaches [4]–[6] and continuous approaches [7]–[11]. Seam-carving based methods are among the most representative discrete approaches [4]–[6]. Based on an energy function, such methods repeatedly remove a spatio-temporal surface until reaching the desired video resolution. However, with complex camera and object motions, finding a surface that does not disturb important video content becomes difficult.

Warping based methods [7]–[11] resize each video frame by finding the optimal warping function of each patch in a continuous domain. For example, Wolf *et al.* [7] formulated video retargeting as solving a least squares problem with sparse linear system equations. As a result, each pixel of low importance is mapped to be relatively close to its neighboring pixels, and vice versa. A few approaches have been proposed to address the temporal incoherence problem which usually happens in video retargeting. To maintain temporal coherence, the method proposed in [8] pre-allocates the space for warping future salient regions by accounting for a shot-time window of succeeding saliency maps. Wang *et al.* [9] proposed to keep temporal consistency with the guide of the optimized motion pathlines of optical flow. The method first performs independent per-frame resizing, then corrects each motion pathline to make its scaling consistent. Consequently, the per-frame resizing is performed again by minimizing the warping error between the independent per-frame resizing and the optimized pathlines. Our previous works [10], [11] proposed to construct a shot-level panoramic mosaic for a video shot to maintain spatio-temporal coherence.

Manuscript received February 01, 2013; revised May 10, 2013; accepted July 08, 2013. Date of publication July 17, 2013; date of current version November 18, 2013. This work was supported in part by the National Science Council, Taiwan, under Grant NSC101-2221-E-007-121-MY3. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Marek Domański.

C.-W. Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

C.-M. Tsai is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 60046, Taiwan.

P.-C. Chen is with the Faraday Technology Corporation, Hsinchu 30013, Taiwan.

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/JSTSP.2013.2273659

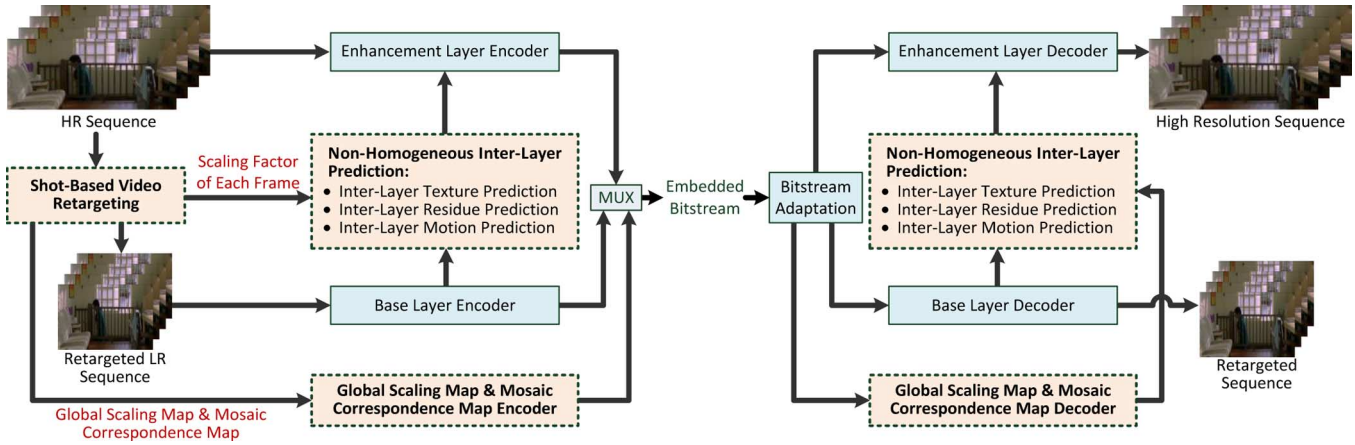

Fig. 1. Block diagram of the proposed content-adaptive spatial scalability coder for SVC.

The panoramic mosaic is used to determine a shot-level global scaling map, which is then used to derive local scaling maps of individual frames after aligning the frames to the mosaic.

To preserve the visual information in the reduced-resolution video in SVC or in applications with bandwidth-limited channel, recently a few content-adaptive coding schemes have been proposed to integrate the content-aware video retargeting with an H.264 coder [13]–[15] or an SVC coder [16]–[18]. Décombas *et al.* [13] proposed to integrate seam carving with H.264/AVC for semantic video coding. In the method, the group of seams and the key lines detection methods are applied to reduce data overhead in storing seam positions. Then, the seam position information is coded for recovering the original resolution video at the decoder side. Then, an improved version of [13] was proposed in [14], in which a new energy map was introduced and the way to define group of seams and the background synthesis method were redesigned. In [15], image epitomes that best represent the original frames are extracted as the priors for efficient intra-coding. The epitomic priors are losslessly coded and sent to the decoder to guide the decoding. Wang *et al.* [16] proposed a content-adaptive spatial scalability framework to extend SVC. A warping based retargeting method [8] was utilized in their framework to adapt the original full-sizes video to content-aware LR video. After that, the warp coding scheme proposed in [17] is used to encode the deformation of each spatial position to its neighboring positions. Finally, three extended inter-layer prediction tools are used to support the non-linear spatial mapping in inter-layer prediction coding. In [18], content-adaptive motion estimation and mode decision schemes were proposed for reducing the computing cost of SVC without sacrificing the coding efficiency.

The non-homogeneous scaling to obtain reduced-resolution video in the above content-adaptive coding schemes, however, requires to send additional side information (e.g., ROI positions, seam positions, warping parameters) to signal the decoder for properly reconstructing the original resolution video in the inter-layer prediction process. The additional side information needs to be efficiently compressed so as to reduce its impact on coding efficiency. However, all the video retargeting methods used in these methods require to save and send side information frame by frame. The per-frame side information consumes a significant amount of memory and channel bandwidth, thereby leading to significant coding efficiency degradation.

To address the problem, we propose a low-overhead content-adaptive spatial-scalability SVC (CASS-SVC) coder which consists of three main modules: a mosaic-guided video retargeter, a side-information coder, and a non-homogeneous inter-layer predictive coder. Instead of sending per-frame side information for the non-homogeneous scaling used in existing methods [13]–[17], without sacrificing the visual quality in the retargeted video, our method only utilizes per-video-shot side information including shot-level global scaling maps and the spatial corresponding positions of individual frames to the panoramic mosaic. Since a video shot usually contains tens of hundreds of video frames, the amount of side information is drastically reduced, thereby significantly increasing the coding efficiency of the SVC. The contribution of the proposed method is three-fold: (i) we propose a new low-overhead CASS-SVC coder; (ii) we propose new shot-based video retargeting schemes that can preserve important visual content as well as maintain spatio-temporal coherence in retargeted video at low computation and overhead costs; and (iii) we propose new non-homogeneous inter-layer prediction tools to achieve good coding efficiency for the proposed CASS-SVC.

The rest of this paper is organized as follows. Section II gives the overview of the proposed CASS-SVC framework. Our shot-based video retargeting method is presented in Section III. The proposed side-information coder is described in Section IV. Section V presents the proposed non-homogeneous inter-layer prediction schemes. Section VI reports the experimental results. Finally, conclusions are drawn in Section VII.

II. Overview of the Proposed Framework

We propose an efficient CASS-SVC coder which aims to preserve important visual information in the LR layer, as well as to reduce the overhead cost of sending side information for guiding the non-homogeneous inter-layer prediction at the decoder without significantly sacrificing the coding efficiency of HR layer. Fig. 1 shows the block diagram of the proposed framework which consists of three modules: (1) the mosaic-guided

TABLE I
NOTATIONS

Symbols	Meanings
W, H	Width and height of the original video
W', H'	Width and height of the resized video
W_G, H_G	Width and height of the global scaling map
$S_x^{(t)}(i,j), S_y^{(t)}(i,j)$	Local scaling factors for the (i,j)-th pixel in the t-th frame
$\mathbf{H}^{(t)}$	Projective transform for the t-th frame
$(i',j')_M$	Projected coordinate of (i,j) in the shot-level panoramic mosaic after frame alignment
$\mathbf{e}_g((i',j')_M)$	Set of energy values corresponding to pixel $(i',j')_M$ of the panoramic mosaic
$(p_L^{(t)}, q_U^{(t)}), (p_R^{(t)}, q_L^{(t)})$	Coordinates of the upper-left and lower-right pixels of the t-th frame in panoramic mosaic
$S_{Gx}^{(n)}((i',j')_M), S_{Gy}^{(n)}((i',j')_M)$	Global scaling factors of $(i',j')_M$ at the n-th round of iterative optimization
D_{info}	Information loss in a resized frame
D_{RI}	Region inconsistency distortion
D_{SFI}	Scaling factor inconsistency distortion
(S_{Gx}^*, S_{Gy}^*)	Optimal global scaling factors
$EL\_BL_x^{(t)}(m,n), EL\_BL_y^{(t)}(m,n)$	EL-BL matrix values for the the (m,n)-th pixel in the t-th HR frame
$pi_x^{(t)}(m,n), pi_y^{(t)}(m,n)$	Phase indexes for the (m,n)-th pixel in the t-th HR frame
$EL\_S_x^{(t)}(p,q), EL\_S_y^{(t)}(p,q)$	Resampling ratios for the (p,q)-th MB in the t-th HR frame

video retargeting module, (2) the side information coder, and (3) the non-homogeneous inter-layer predictive coder.

By constructing shot-level panoramic mosaics, the shot-based video retargeting module non-homogeneously resizes frames in a shot to preserve important visual content in the frames while maintaining intra-frame spatial coherence and inter-frame temporal coherence with the help of the panoramic mosaics. After obtaining the retargeted LR video, the global scaling maps and mosaic correspondence map are respectively coded by the side information coder and sent to the decoder for guiding the reconstruction of the HR video. Conventional SVC coders reduce data redundancy between two adjacent spatial-resolution layers by using linear inter-layer prediction coding tools which subtract the up-scaled co-located blocks of LR layer from the HR counterpart, and code and then embed the prediction residues into the scalable bitstream. Because the corresponding spatial relationship of two adjacent spatial layers in the proposed framework is no longer linear, we design a mapping matrix to rebuild the spatial correspondences of each frame. With the guide of the mapping matrix, we redesign a non-homogeneous up-scaling operation of interlayer prediction coding tools to achieve good prediction accuracy.

III. SHOT-BASED VIDEO RETARGETING

Table I lists the main symbols used in this paper. Assume we resize a video from resolution $W \times H$ to $W' \times H'$. Content-adaptive video retargeting aims to determine the scaling factors of individual pixels in each video frame based on their content importance so as to maximize in the downsized video the retained information and perceptual quality subject to a predetermined size budget. Let $S_x^{(t)}(i,j)$ and $S_y^{(t)}(i,j)$ respectively denote the horizontal and vertical scaling factors for the (i,j)-th pixel in the t-th frame. The sums of scaling factors in each row

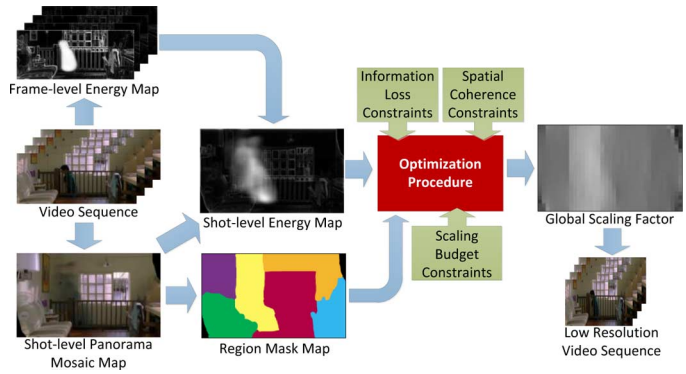

Fig. 2. Block diagram of the proposed shot-based video retargeting method.

and in each column are respectively given as $\sum_{i=1}^{W} S_x^{(t)}(i,j) = W',\ \forall j$ and $\sum_{i=1}^{W} S_x^{(t)}(i,j) = H',\ \forall j$.

Since a CASS-SVC bitstream consumes an additional bitrate for sending the side information to guide non-homogeneous resizing, the amount of side information plays a crucial role in coding efficiency. The video retargeting scheme used in a CASS-SVC should not only preserve as much information as possible in the downsized video, but also keeps the consumed overhead minimal. To this end, we propose a low-overhead shot-based video retargeting method by modifying our previous works [10],[11] to significantly reduce the amount of side information while still maintaining good visual quality of retargeted video. Besides, our approach addresses the spatio-temporal incoherence problem commonly faced in video retargeting in a systematic and efficient way, rather than resorting to numerous per-frame optimizations[9], [10].

As illustrated in Fig. 2, similar to our previous work[10], the proposed method first performs shot boundary detection [19] and constructs a panoramic mosaic for each video shot. The mosaic image is then segmented into regions by using a semi-automatic segmentation tool proposed in [20]. Before deriving the scaling maps of individual frames, our method first retargets the shot-level panoramic mosaic to the desired scaling ratio. Based on the shot-level panoramic mosaic, the region segmentation masks and a set of spatial coherence constraints, information loss constraints, and scaling budget constraints, we propose an iterative optimization scheme to obtain a shot-level global scaling map for the mosaic, which is then used to derive frame-level scaling maps at both the encoder and decoder.

The proposed retargeting scheme is a coding-friendly version of our previous method proposed in[11]. Unlike the scheme [10] which requires to perform per-frame optimization to derive frame-level scaling maps, we embed the per-frame scaling budgets into the constraints for the iterative optimization of global scaling map. As a result, frame-level scaling maps can be directly computed from the global scaling map at both the encoder and decoder without resorting to the per-frame optimization. This does not only reduce the computation cost but also drastically reduces the amount of side information since only the global map along with a few frame alignment information need to be sent to the decoder as side information. Note, sending a coded shot-level global scaling map of size $W_G \times H_G$ (the size is about the same order as a frame size, see Table II in

TABLE II
COMPLEXITY OF THE TWO PROPOSED RETARGETING METHODS

Movie Title	Die Hard 4.0	White Car	X Man
Number of Frames	181	87	167
Panorama Mosaic Size	807×290	1190×390	981×286
Method in (9)	815.8 sec	1329.8 sec	839.8 sec
Method in (12)	187.2 sec	66.0 sec	45.3 sec

Section VI) instead of sending numerous (usually tens or hundreds) frame-level local scaling maps of size $W \times H$ would significantly reduce the overhead bitrate. To further reduce the bitrate of the global scaling map, we impose an additional spatial constraint to limit the same column/row of the shot-level panoramic mosaic to share the same scaling factor. As a result, the number of the global scaling factors is reduced from $W_G \times H_G$ to $W_G + H_G$. After calculating the frame-level scaling maps based on the shot-level global scaling map, we perform a separable 2D retargeting: a horizontal resizing followed by a vertical resizing. The detailed operations of our retargeting scheme are elaborated below.

A. Initialization

At the encoder, the proposed video retargeting method exploits four kinds of information to resize a video shot: the frame-level energy maps, the shot-level panoramic mosaic, the shot-level energy map, and the region segmentation mask.

1) Frame-Level Energy Maps: The energy function, which is used to represent the pixel-wise visual importance in each video frame, plays an important role in content-aware video retargeting. Using the saliency detection model proposed in [12], which is based on the human visual sensitivity and the amplitude spectrum of quaternion Fourier transform, we obtain each frame's energy (saliency) map $e(i,j)$ which represents the energy value of the (i,j)-th pixel in a frame.

2) Shot-Level Panoramic Mosaic: The shot-level panoramic mosaic is generated to provide a global reference of the positions of video frames in a shot. Similar to our previous work[10], our method uses the SIFT [19] descriptor to select feature points in each video frame and to perform correspondence matching between two neighboring frames. We use a simplified affine model with only scaling and translation parameters. Furthermore, we use RANSAC [22], [23] to estimate the camera motion between two neighboring frames based on the correspondences matched by SIFT. Consequently, the panoramic mosaic of a shot is generated by using the estimated camera motion parameters to align the frames in a shot.

3) Shot-Level Energy Map: The shot-level energy map is obtained by fusing all frame-level energy maps of a shot based on the corresponding locations in the shot-level panoramic mosaic.

Let $\mathrm{H}^{(t)}$ denote the projective transform for the t-th frame, (i,j) the coordinate of the (i,j)-th pixel in the t-th frame, and $(i',j')_M$ the projected coordinate of (i,j) in the shot-level panoramic mosaic after frame alignment. Then, the projection of a coordinate is represented by $(i',j')_M = \mathbf{H}^{(t)}(i.j)$. The shot-level energy map is simply obtained as the mean of the energy values corresponding to the same pixel after the frame alignment, as expressed by

$$e_G((i',j')_M) = \mathrm{mean}\left\{ \bigcup_t \bigcup_{(i,j)\to(i',j')_M} e\left(\mathbf{H}^{(t)}(i,j)\right) \right\} \tag{1}$$

where $\mathrm{e}_G\left((i',j')_M\right)$ denote the energy value corresponding to pixel $(i',j')_M$ of the panoramic mosaic.

Note, the union operation in (1) is a many-to-one mapping, that is, a pixel $(i',j')_M$ in the panoramic mosaic may be associated with the energy values from multiple frame-level energy maps. To obtain a single-valued mapping, we use the mean value of the set in (1) as the energy value for pixel $(i',j')_M$ of the shot-level energy map.

4) Panoramic Mosaic Segmentation Mask: The scaling factors of pixels/patches inside each region should be kept as consistent as possible. To this end, we use the semi-automatic object segmentation tool proposed in [20] to identify objects/regions in the panoramic mosaic for each shot. Note, the segmentation result by the automatic segmentation tool proposed in [20] still has over-segmentation problem. Therefore, the user can scribble the segmentation image to merge the regions that belong to the same object.

B. Scaling Budget Constraints

A key step of the proposed method is to embed the per-frame scaling budgets for a video shot in the constraints imposed in the iterative optimization process to derive the shot-level global scaling map. This can achieve fairly good accuracy while directly computing the per-frame local scaling maps from the global scaling map at both the encoder and decoder, without the need of resorting to computationally expensive per-frame optimization and the cost of transmitting bandwidth demanding local scaling maps. For the shot-level panoramic mosaic, its available scaling budget is constrained by the corresponding frame alignment location and the target resolution.

Suppose the corresponding coordinates of the upper-left and lower-right pixels of the t-th frame in the panoramic mosaic are $\left(p_L^{(t)}, q_U^{(t)}\right)$ and $\left(p_R^{(t)}, q_L^{(t)}\right)$, respectively, i.e., $\left(p_L^{(t)}, q_U^{(t)}\right) = \mathrm{H}^{(t)}(1,1)$ and $\left(p_R^{(t)}, q_L^{(t)}\right) = \mathrm{H}^{(t)}(W,H)$. Recall that we use a simplified affine model with only scaling and translation parameters, the horizontal and vertical scaling budget constraints for the t-th frame are set as

$$\begin{cases} \sum_{i'=cp_L^{(t)}}^{p_R^{(t)}} S_{Gx}^{(n)}\left((i',j')_M\right) = \left(p_R^{(t)} - p_L^{(t)} + 1\right) \times \left(\frac{W'}{W}\right) \\ \sum_{j'=q_u^{(t)}}^{q_L^{(t)}} S_{Gy}^{(n)}\left((i',j')_M\right) = \left(q_L^{(t)} - q_U^{(t)} + 1\right) \times \left(\frac{H'}{H}\right) \end{cases}$$
$$\forall q_U^{(t)} \le j' \le q_L^{(t)} \mathrm{and} \forall p_L^{(t)} \le i' \le p_R^{(t)} \tag{2}$$

where $S_{Gx}^{(n)}\left((i',j')_M\right)$ and $S_{Gy}^{(n)}\left((i',j')_M\right)$ denote the horizontal and vertical global scaling factors of $(i',j')_M$ at the n-th round of iterative optimization. Note, for downscaling, $0 \le S_{Gx}^{(n)}\left((i',j')_M\right) \le 1$ and $0 \le S_{Gy}^{(n)}\left((i',j')_M\right) \le 1$.

With the constraints in (2), we can compute the scaling map of a video frame with a satisfactory accuracy directly from the

panoramic mosaic according to the available scaling budget without resorting to per-frame optimization procedure in [9] and [10]. Even if the optimization procedure may determine improper scaling factors for a certain region, the error propagation is only limited to the frames containing the region.

C. Information Loss Constraints

The information loss after resizing a panoramic mosaic can be measured by the sum of the products of pixel-wise energy value and subsampling ratios in the panoramic mosaic in the x and y directions as follows:

$$\begin{aligned} D_{\text{info}} &= D_{\text{info}}\left(S_{Gx}^{(n)}\right) + D_{\text{info}}\left(S_{Gy}^{(n)}\right) \\ &= \sum_{i',j'}\left(1 - S_{Gx}^{(n)}\left((i',j')_M\right)\right)\Delta e_G\left((i',j')_M\right) \\ &\quad + \sum_{i',j'}\left(1 - S_{Gy}^{(n)}\left((i',j')_M\right)\right)\Delta e_G\left((i',j')_M\right). \end{aligned} \quad (3)$$

D. Spatial Coherence Constraints

In the optimization process, we impose the following costs and constraints to avoid the spatial incoherence distortion.

1) Cost of Region Inconsistency Distortion: To maintain spatial coherence, the scaling factors within a region should be made consistent. To do so, we define a set $R = \{R_1, R_2, \ldots, R_K\}$ consisting of all regions in the region segmentation mask of the shot-level panoramic mosaic, where K is the number of labeled regions. To maintain the consistency of pixels' resizing ratios in a region, the region inconsistency distortion of the k-th region is first measured by

$$\begin{aligned} D_{\text{RI}}(R_k) = \left(\text{std}\left(S_{Gx}^{(n)}((i',j')_M)\right) + \text{std}\left(S_{Gy}^{(n)}((i',j')_M)\right)\right) \\ \cdot e_G\left((i',j')_M\right),\ \forall (i',j')_M \in R_K, \end{aligned} \quad (4)$$

where $\text{std}(\cdot)$ denotes the standard deviation function. The region inconsistency distortion of all objects can be obtained by

$$D_{\text{RI}} = \sum_{k=1}^{K} D_{\text{RI}}(R_k). \quad (5)$$

In brief, we minimize the variation of scaling factors in each region to maintain the consistency of each region's size.

2) Spatial Smoothness Constraints: To reduce the bit-rate of side information, we impose the following spatial smoothness constraints:

$$\begin{cases} S_{Gx}^{(n)}\left((i',j')_M\right) = S_{Gx}^{(n)}\left((i',j'+1)_M\right),\ \forall i' \\ S_{Gy}^{(n)}\left((i',j')_M\right) = S_{Gy}^{(n)}\left((i'+1,j')_M\right),\ \forall j' \end{cases}. \quad (6)$$

In other words, when performing horizontal resizing, the pixels within the same column/row share the same scaling factor value. This reduces the number of global scaling factors for the panoramic mosaic from $W_G \times H_G$ to $W_G + H_G$, thereby reducing the bitrate for coding the global scaling map. Besides, to control the scaling factor changes between adjacent columns/rows, we introduce additional constraints into the optimization proceduree as follows:

$$\begin{cases} |S_{Gx}^{(n)}\left((i',j')_M\right) - S_{Gx}^{(n)}\left((i'+1,j')_M\right)| \le TH_{Sx},\ \forall i\prime \\ |S_{Gy}^{(n)}\left((i',j')_M\right) - S_{Gy}^{(n)}\left((i',j'+1)_M\right)| \le TH_{Sy},\ \forall j' \end{cases} \quad (7)$$

where TH_{Sx} and TH_{Sy} are used to control the deformation degree. Larger TH_{Sx} and TH_{Sy} would preserve more important visual content in the retargeted video, but would also raise the risk of introducing unacceptable structure deformation artifacts. We set TH_{Sx} and TH_{Sy} to 0.001 in all our experiments.

3) Cost of Scaling Inconsistency Distortion: To avoid aspect ratio deformation, the horizontal and vertical scaling factors of visually important pixels should also be kept consistent. Therefore, as part of the cost, we measure the scaling inconsistency distortion, which is formulated as the sum of the absolute differences between a pixel's horizontal and vertical scaling factors weighted by the pixel's energy value:

$$\begin{aligned} &D_{\text{SI}} \\ &= \sum_{i',j'}\left\{|S_{Gx}^{(n)}\left((i',j')_M\right) - S_{Gy}^{(n)}\left((i',j')_M\right)|\Delta e_G\left((i',j')_M\right)\right\}. \end{aligned} \quad (8)$$

E. Iterative Optimization for the Shot-Level Scaling Map

An iterative optimization procedure is performed to find a converged solution $\left(S_{Gx}^*,\ S_{Gy}^*\right)$ by minimizing the overall distortion involving (3), (5), (8) subject to the scaling budget and spatial smoothness constraints. The optimization procedure is formulated as a nonlinear optimization problem of the form:

$$\left(S_{Gx}^*,\ S_{Gy}^*\right) = \underset{\left(S_{Gx}^{(n)},\ S_{Gy}^{(n)}\right)}{\arg\min}\left(D_{\text{info}} + \lambda_1 D_{\text{RI}} + \lambda_2 D_{\text{SI}}\right), \quad (9)$$

where $\left(S_{Gx}^*,\ S_{Gy}^*\right)$ are the optimal global scaling factors in the x and y directions. We use the interior-point solver [24] to solve the nonlinear optimization problem. The weights for D_{R1} and D_{S1} are set equal (i.e., $\lambda_1 = \lambda_1 = 1$). Suppose the height and width of the shot-level panoramic mosaic map are H_G and W_G, then the initial scaling factor values are set as follows:

$$\begin{cases} S_{Gx}^{(1)}\left((i',j')_M\right) = \frac{\sum_{y=1}^{H_G} e_G\left((i',y)_M\right)}{H_G} \\ S_{Gy}^{(1)}\left((i',j')_M\right) = \frac{\sum_{x=1}^{W_G} e_G\left((x,j')_M\right)}{W_G} \end{cases}. \quad (10)$$

F. Computation of Frame-Level Scaling Maps

After obtaining the optimal global scaling factors, $\left(S_{Gx}^*,\ S_{Gy}^*\right)$, the local scaling map of the t-th frame is computed via the projective transform by

$$\begin{cases} S_x^{(t)}(i,j) = S_{Gx}^*\left(\text{H}^{(t)}(i,j)\right) = S_{Gx}^*(i',j')_M \\ S_y^{(t)}(i,\ j) = S_{Gy}^*\left(\text{H}^{(t)}(i,j)\right) = S_{Gy}^*(i',j')_M \end{cases}. \quad (11)$$

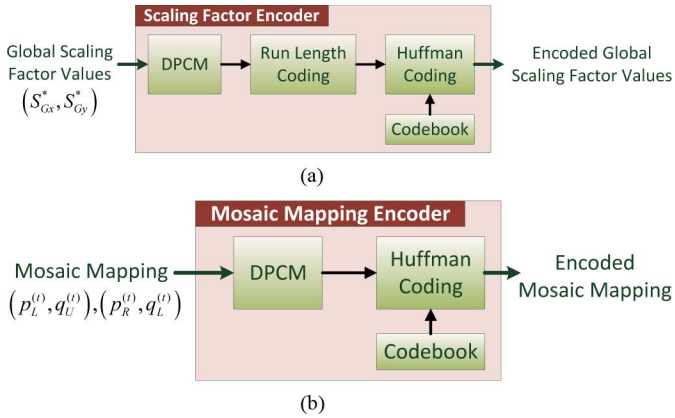

Fig. 3. Proposed side-information coders: (a) the scaling map coder; (b) the mosaic correspondence map coder.

After obtaining the local scaling factors, the LR frames are generated by the pixel fusion based image downscaling proposed in [7].

G. Low-Cost Mosaic-Guided Video Retargeting

Since the nonlinear optimization method in (9) is still time consuming, a low-cost simplification is to relax the D_{R1} and D_{S1} of the cost function in (9) as follows:

$$\left(S^*_{Gx},\ S^*_{Gy}\right) = \arg\min_{\left(S^{(n)}_{Gx},\ S^{(n)}_{Gy}\right)} D_{\mathrm{info}} \quad \text{subject to (2), (6) and (7).} \tag{12}$$

As a result, the optimal global scaling map can be obtained by using a linear programming solver [25], which significantly reduces computation in determining the global scaling factors. However, due to the removal of the spatial coherence costs, the resized images might have a few inconsistent deformation and noticeable visual artifacts.

IV. Side-Information Coders

SVC exploits interlayer prediction and coding tools to enhance the coding efficiency for spatial scalability. The "I_BL" type macroblock (MB) is coded by a spatial scalability coding tool by which the HR block is reconstructed by adding prediction residues to the corresponding up-scaled LR block. Since LR frames in CASS-SVC are non-homogeneously downscaled from HR frames, additional side information is needed to signal the decoder to up-scale a LR frame correctly. The global scaling map [i.e., $\left(S^*_{Gx},\ S^*_{Gy}\right)$] and the correspondence map of individual frames to the panoramic mosaic [i.e., $\left(p_L^{(t)}, q_U^{(t)}\right)$ and $\left(p_R^{(t)}, q_L^{(t)}\right)$] are the side information required for correctly up-scaling an LR frames to an HR frames at both the encoder and decoder. Therefore, we design two coders to encode the global scaling map and the correspondence map derived from the panoramic mosaic.

A. Global Scaling Map Encoder

Fig. 3(a) shows the block diagram of the global scaling map encoder that uses DPCM, run-length coding (RLC), and Huffman coding. The horizontal and vertical global scaling factors, S^*_{Gx} and S^*_{Gy}, are separately encoded. Due to the imposed spatial constraints in (6) and (7), the pixels in a region of the panoramic mosaic would have similar scaling factor values, where the variations of adjacent scaling factors are controlled within TH_{Sx} and TH_{Sy}. Since adjacent scaling factor values are close, DPCM is applied to remove the spatial redundancy among the values, followed by RLC to encode the nonzero prediction residues. As a result, the global scaling map is compactly represented by a sequence of 2-D number pairs in the form of (`RUNLENGTH`, `VALUE`), where `VALUE` denotes the differential value after DPCM, and `RUNLENGTH` denotes the number of consecutive zeros between two nonzero `VALUE`. Finally, Huffman coding is used to remove the statistical redundancy among these 2D symbols. To design the Huffman coder, we resize various types of video clips to collect the scaling maps as training data. To encode the scaling factors, DPCM is applied and the numbers of consecutive zeros between nonzero scaling factors are also recorded. Then, the statistics of differential mosaic correspondence values are calculated and the Huffman code is assigned according to the statistics. Note, in our implementation, the precision of scaling factors is set to 10^{-5} for a reasonable size of codebook.

B. Mosaic Correspondence Map Encoder

As shown in Fig. 3(b), DPCM and Huffman coding are used to encode the correspondence map of individual frames to the shot-level panoramic mosaic. The four correspondence values, $\left(p_L^{(t)}, q_U^{(t)}\right)$ and $\left(p_R^{(t)}, q_L^{(t)}\right)$, are separately encoded. Unlike the global scaling map coder, since the camera motion is unpredictable, RLC is not suitable to encode the correspondence values. After being coded by DPCM, the residues are directly encoded by using Huffman coding.

V. Non-Homogeneous Interlayer Predictive Coders

We modified the spatial scalability coding tools in SVC to support the non-homogeneous inter-layer prediction used in CASS-SVC. Four re-designed coding tools are used: 1) the EL-BL mapping matrix, which records the spatial correspondences of each frame from the HR layer (i.e., the enhancement-layer) to the LR layer (i.e., the base-layer), 2) the non-homogeneous inter-layer texture prediction, 3) the non-homogeneous inter-layer residue prediction, and 4) the non-homogeneous inter-layer motion prediction.

A. The EL-BL Mapping Matrix

In our method, the EL-BL mapping matrix is designed to indicate the spatial correspondences of each frame from an HR frame to its retargeted LR frame. Because SVC only supports frame cropping and uniform scaling to adapt video resolutions or aspect ratio, a linear mapping function is used in SVC to derive the spatial correspondences between two adjacent spatial layers[2]. Nevertheless, since the mapping function in CASS-SVC is no longer linear, the EL-BL mapping matrix is used to indicate the nonlinear mapping relationships.

A general case of creating the EL-BL mapping matrix of each frame is given here. The correspondences from an HR frame to an LR frame is derived from the local scaling factors, $S_x^{(t)}$

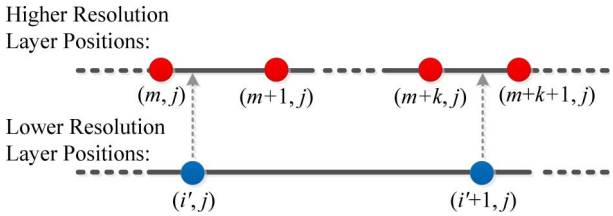

Fig. 4. Illustration of non-linear spatial mapping between two adjacent spatial layers in the horizontal direction.

and $S_y^{(t)}$. Fig. 4 illustrates the nonlinear spatial mapping between two adjacent spatial layers in the horizontal direction. After horizontal warping, suppose the horizontal positions of the $(m+1, j)$-th to $(m+k, j)$-th pixels in the HR frame correspond to the fractional positions between (i', j) and $(i'+1, j)$ in the LR frame. Then, the corresponding EL-BL matrix values of the horizontal positions from $(m+1, j)$ to $(m+k, j)$ are all set to i'. The EL-BL matrix values of the t-th frame in x and y directions are computed by

$$\begin{cases} EL\_BL_x^{(t)}(m,n) = \text{Floor}\left(\sum_{i=1}^{m} S_x^{(t)}(i,n)\right) + 1, \forall n \\ EL\_BL_y^{(t)}(m,n) = \text{Floor}\left(\sum_{j=1}^{n} S_y^{(t)}(m,j)\right) + 1, \forall m \end{cases} \tag{13}$$

where $\text{Floor}(\cdot)$ stands for the floor function which takes the largest integer smaller than the input.

B. Non-Homogeneous Interlayer Texture Prediction

In the interlayer texture prediction in SVC, if a MB of the HR layer is coded as the I_BL MB type, the co-located blocks in the LR layer is up-scaled and subtracted from the corresponding MB of the HR layer. The prediction residue is then intra coded. When up-scaling the blocks in the LR layer, the luma component is up-scaled by a separable 4-tap polyphase interpolation filter, and chroma components are up-scaled by a bilinear interpolation filter [2]. Note that, a phase index is used to indicate the filter coefficients according to the spatial position of up-scaling when applying the interpolation filters. In our method, the up-scaling operation is modified to support the non-homogeneous prediction as shown in Fig. 5(a), where the phase index is determined according to the scaling factor, and the corresponding pixel values used in the interpolation are located by the EL-BL mapping matrix of each frame.

The phase indexes of the t-th frame are determined by

$$\begin{cases} pi_x^{(t)}(m,n) = \mathcal{F}\left(\sum_{i=1}^{m} S_x^{(t)}(i,n)\right) \times 16 \\ pi_y^{(t)}(m,n) = \mathcal{F}\left(\sum_{j=1}^{n} S_y^{(t)}(m,j)\right) \times 16 \end{cases} \tag{14}$$

where $\mathcal{F}(\cdot)$ is a function for obtaining the fractional part of the input value, that is, $\mathcal{F}(A) = A - \text{Floor}(A)$.

The four spatial positions used in the 4-tap polyphase interpolation filter are derived from the EL-BL mapping matrix. While performing the horizontal interposition at the (m,n)-th pixel, the four spatial positions are $\left(EL\_BL_x^{(t)}(m,n) - 1, EL\_BL_y^{(t)}(m,n)\right)$, $\left(EL\_BL_x^{(t)}(m,n), EL\_BL_y^{(t)}(m,n)\right)$, $\left(EL\_BL_x^{(t)}(m,n) + 1, EL\_BL_y^{(t)}(m,n)\right)$, and $\left(EL\_BL_x^{(t)}(m,n) + 2, EL\_BL_y^{(t)}(m,n)\right)$, respectively. Note that, due to 4:2:0 color sub-sampling, both the width and height of chroma components are only half of those of luma component. Therefore, a 13-taps dyadic interpolation filter [26] is applied to (S_{Gx}^*, S_{Gy}^*) to generate the scaling factors for chroma components. The phase indexes of chroma components are derived in the same manner, as formulated in (14).

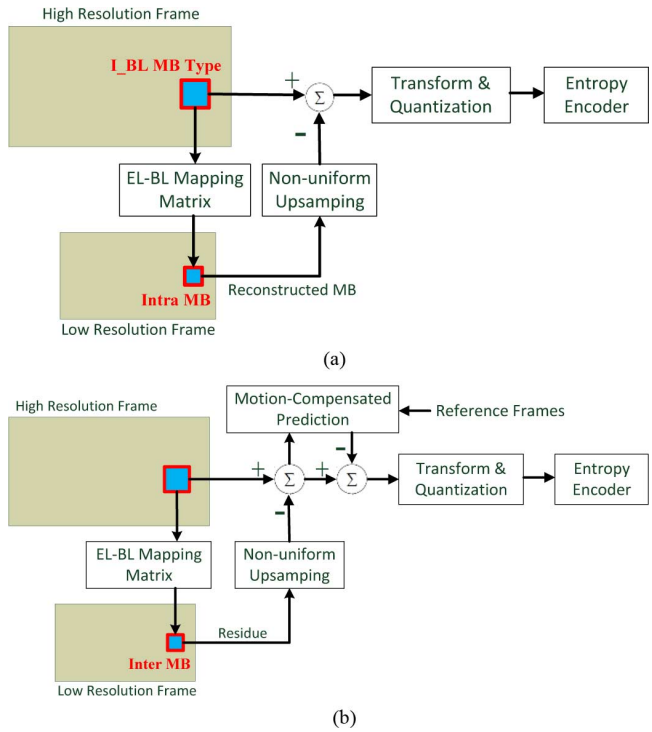

Fig. 5. Proposed non-homogeneous inter-layer predictive coders: (a) the texture prediction coder; (b) the residue prediction coder.

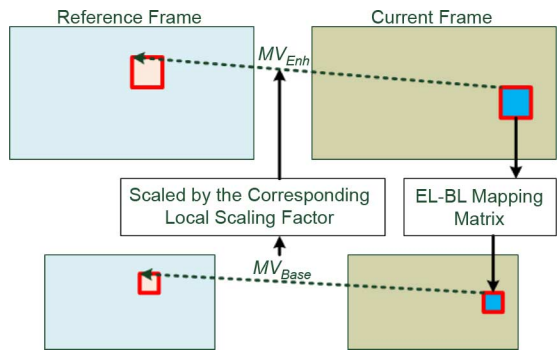

Fig. 6. Proposed inter-layer motion prediction coding structure.

C. Non-Homogeneous Interlayer Residue Prediction

Similar to the method described in Section VI-B, as shown in Fig. 5(b), we redesign the interlayer residual prediction scheme to support the non-homogeneous interlayer prediction in CASS-SVC. The interlayer residue prediction in SVC is performed when the corresponding spatial position in the LR layer is inter-coded. If the residue prediction is activated, the motion-compensated prediction residues of the co-located blocks in the LR layer is up-scaled and subtracted from the residues of the corresponding MB in the HR layer. The differences of residues are

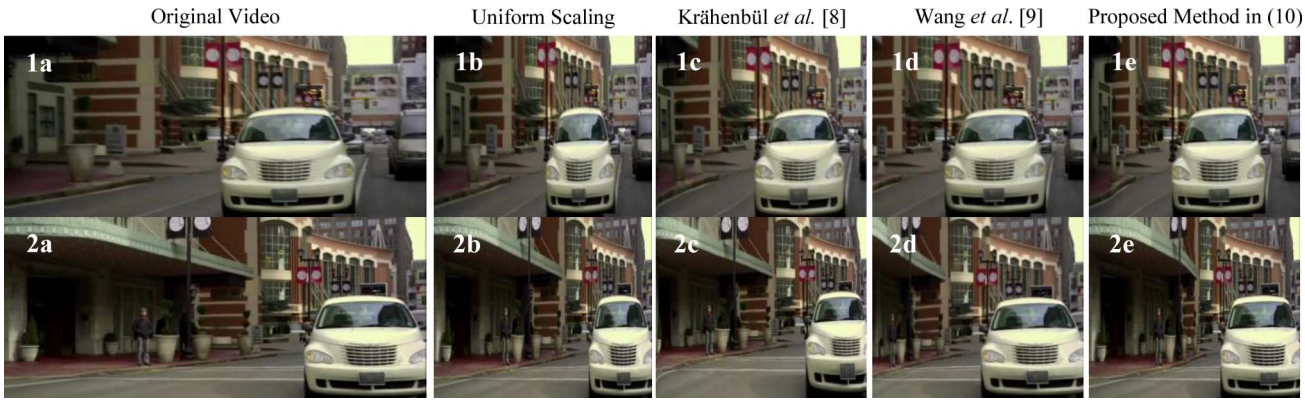

Fig. 7. Subjective quality comparison of the proposed method with uniform scaling, the retargeting scheme proposed by Krähenbühl *et al.* [8], and the retargeting scheme without cropping operator proposed by Wang *et al.* [9].

then coded and embedded into the scalable bitstream. When upscaling the blocks in the LR layer, luma and chroma components are all upscaled by a bilinear interpolation filter [2]. Similarly, a phase index is used to indicate the interpolation filter coefficients. In our method, the upscaling operation is also modified to support the non-homogeneous prediction relations. The phase index decision method is the same as (15). Moreover, the spatial positions used in the filtering operation cannot be outside the block boundary; otherwise, it will cause coding artifact. Similarly, the chroma components are downscaled by the factor of two using a 13-tap interpolation filter [26].

D. Non-Homogeneous Interlayer Motion Prediction

Interlayer motion prediction is another coding tool used in SVC to reduce the bitrate of motion data in the HR layer. In SVC, when the motion prediction mode is activated, the motion vector of a MB in the HR layer is predicted either from the neighboring MBs of same layer or from the upscaled motion vector of the corresponding LR-layer MB. To obtain correct resampling ratios between two adjacent spatial layers for non-homogeneous upscaling, as shown in Fig. 6, our method first uses the EL-BL mapping matrix to retrieve the corresponding LR-layer motion vectors. The resampling ratio for the (p, q)-th MB is then computed by (15). Then, the corresponding LR-layer motion vectors are upscaled by the resampling ratio.

$$\begin{cases} EL\_S_x^{(t)}(p,q) = \frac{16}{\sum_{i=16\times p+1}^{16*p+16}} S_x^{(t)}(i, q\times 16+1) \\ EL\_S_y^{(t)}(p,q) = \frac{16}{\sum_{j=16\times q+1}^{16*q+16}} S_y^{(t)}(p\times 16+1, j) \end{cases}. \quad (15)$$

VI. Experiments and Discussion

In this section, we compare the performances of the proposed video retargeting scheme and the proposed CASS-SVC with other exiting methods. For subjective evaluation, readers can obtain the complete set of test results from our project website [28].

A. Performance Evaluation of Shot-Based Video Retargeting

To evaluate the performance of the proposed video retargeting scheme, we select test sequences which contain rich types of camera and object motions from cinema and drama videos. In our experimental setting, each test video is resized to the half size of the original width. We compare our method with three exiting schemes including the uniform scaling and the two state-of-the-art warping-based retargeting schemes proposed in [8] and [9], respectively. To make a fair comparison, the cropping operator of the method in[9] is disabled.

Fig. 8 shows the subjective quality comparisons of the proposed method and the other three methods. Obviously, uniform scaling, which is used in conventional SVC, is immune to spatio-temporal incoherence distortion caused by camera or object motions. It, however, results in small sized objects and background in important regions and the change on an object's aspect ratio. Due to ignoring the temporal corresponding relationships and the limited windows size, the method proposed in [9] may cause some visual artifacts, such as stretching, shrinking, or both, on important video content, and obvious geometric deformations around frame boundary regions. For example, Figs. 7(1c) and 7(2c) illustrate the car sizes are inconsistent, and Fig. 7(2c) shows obvious geometric deformations around frame boundaries. Wang *et al.* [9] proposed to keep temporal consistency with the guide of the optimized motion pathlines of optical flow. Because the detected motion pathline inside the newly appearing and disappearing area may be unreliable (e.g., in the frame boundary regions), the final motion-guided resizing step would be dominated by other motion pathlines to keep temporal consistency. However, if the saliency of these unreliable motion pathlines is unapparent, these areas may be over-squeezed as can be observed in Figs. 7(1d) and 7(2d). In contrast, with the guide of shot-based mosaic, the proposed method can maintain the spatio-temporal coherence as well as preserve the content structure in each frame by evaluating the global scaling map, so as to avoid such visual artifacts.

To compare our retargeting method based on (9) and its low-cost version based on (12), we also compare the scaling maps obtained by our previous work[11], the optimization methods in (9) and (12). Fig. 8(a) to Fig. 8(c) illustrate the panorama mosaic image, the shot-level energy map and the object segmentation mask, respectively. Figs. 8(d2)–8(d3) show that our previous work[11] successfully maintains coherent object sizes. As can be observed from the snapshots depicted

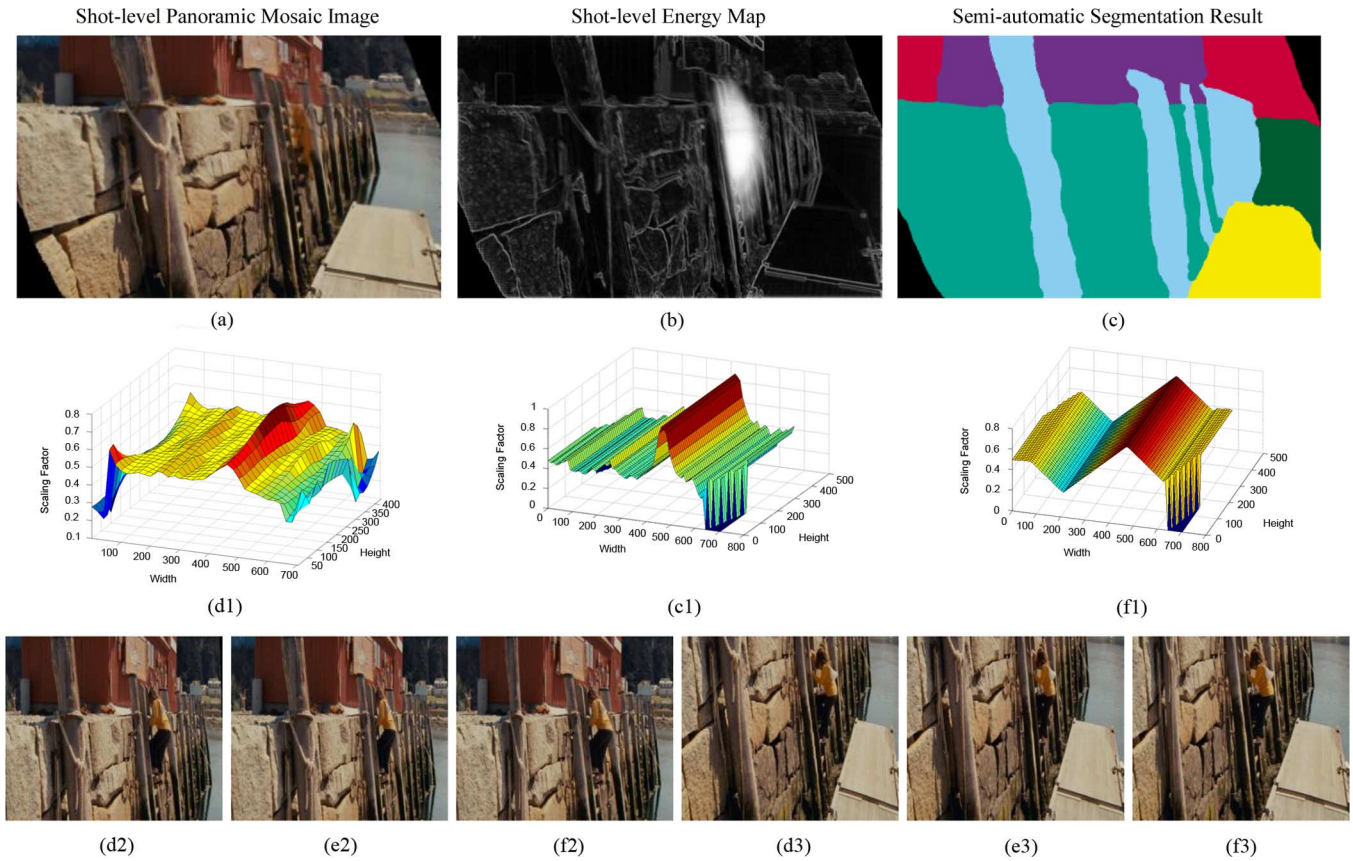

Fig. 8. Comparison of the obtained global scaling maps: (a) Shot-level panorama mosaic image; (b) Shot-level energy map; (c) Object segmentation mask; (d1)(e1)(f1) The global scaling maps derived by using different optimization methods in [11], (10) and (12), respectively; (d2)-(d3) Snapshots obtained by the per-frame optimization method in [11]; (e2)-(e3) Snapshots obtained by the method in (10); (f2)-(f5) Snapshots obtained by the method in (12).

in Figs. 8(e2)–8(e3), the subjective visual quality achieved by the optimization method in (9) is comparable with that of our previous method [10], with a slight difference in the preservation of salient regions due to the spatial smoothness constraints imposed in deriving the global scaling map. Nevertheless, our method in (9) leads to much fewer side information and low computational complexity, which is particularly good for video coding. Figs. 8(f2)–8(f3) show the results of using the low-cost optimization method in (12) that only considers the information loss cost. Compared with the other two results, the salient regions are still preserved satisfactorily, but some background regions are over-squeezed. In summary, our methods derive the frame-level scaling maps by directly retargeting the panoramic mosaic image, making the computation cost and overhead cost much lower while still maintaining satisfactory visual quality. Table II lists the time consumption of evaluating the scaling factors of three videos. Our method was implemented using Matlab® on a personal computer with Intel Core i5 2430 M CPU and 6 GB memory. Note, the methods proposed in [8] and [10] require to solve an optimization problem with $\mathrm{O}(W \times H \times T)$ unknown variables for a video shot or a time window, where T is the shot length or the size of time-window. The method in [9] requires to solve $\mathrm{O}(T)$ individual optimizations, each having $\mathrm{O}(W \times H)$ unknow variables. The proposed method only requires to solve one optimization problem with $\mathrm{O}(W_G + H_G)$ unknown variables for a video shot.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED CASS-SVC AND SVC USING THE TWO QP SETTINGS

Sequence	Size	(QP^1_{BL}, QP^1_{EL})			(QP^2_{BL}, QP^2_{EL})		
		BDBR	BDPSNR	Overhead	BDBR	BDPSNR	Overhead
Alvin	640x336	3.56%	-0.12dB	0.18%	3.21%	-0.15dB	0.12%
How	656x368	6.26%	-0.24dB	0.05%	3.43%	-0.25dB	0.03%
Concert	688x288	2.87%	-0.06dB	0.04%	2.60%	-0.08dB	0.02%
Upstairs	688x288	4.77%	-0.15dB	0.10%	5.87%	-0.25dB	0.07%
Angry	688x288	6.79%	-0.24dB	0.31%	4.93%	-0.22dB	0.22%
Die Hard 4.0	672x288	2.82%	-0.08dB	0.05%	3.79%	-0.14dB	0.04%
X Men 2	608x256	5.27%	-0.11dB	0.14%	5.37%	-0.14dB	0.09%
ParkScene	1280x720	3.46%	-0.13dB	0.01%	3.97%	-0.24dB	0.01%
Kimono	1280x720	9.51%	-0.36dB	0.02%	5.45%	-0.37dB	0.01%
Ducks take off	1920x1080	3.17%	-0.08dB	0.0003%	3%	-0.07dB	0.0001%
Sunflower	1920x1080	6.37%	-0.15dB	0.01%	3.44%	-0.37dB	0.007%
Traffic	2560x1440	4.5%	-0.11dB	0.001%	2.78%	-0.01dB	0.0003%
Average		**4.91%**	**-0.16 dB**	**0.004%**	**3.92%**	**-0.16dB**	**0.004%**

B. Coding Performance of CASS-SVC

To evaluate the coding performance of the proposed CASS-SVC, we compare the rate-distortion performance of our method with the SVC and the method proposed in [15]. The test videos are of various spatial resolutions, including high definition (HD) and non-HD sequences as listed in Table III. In the experimental settings, the width of each test sequence is halved (e.g., from 688 × 288 to 384 × 288). Each test sequence is coded using the hierarchical B-picture prediction structure

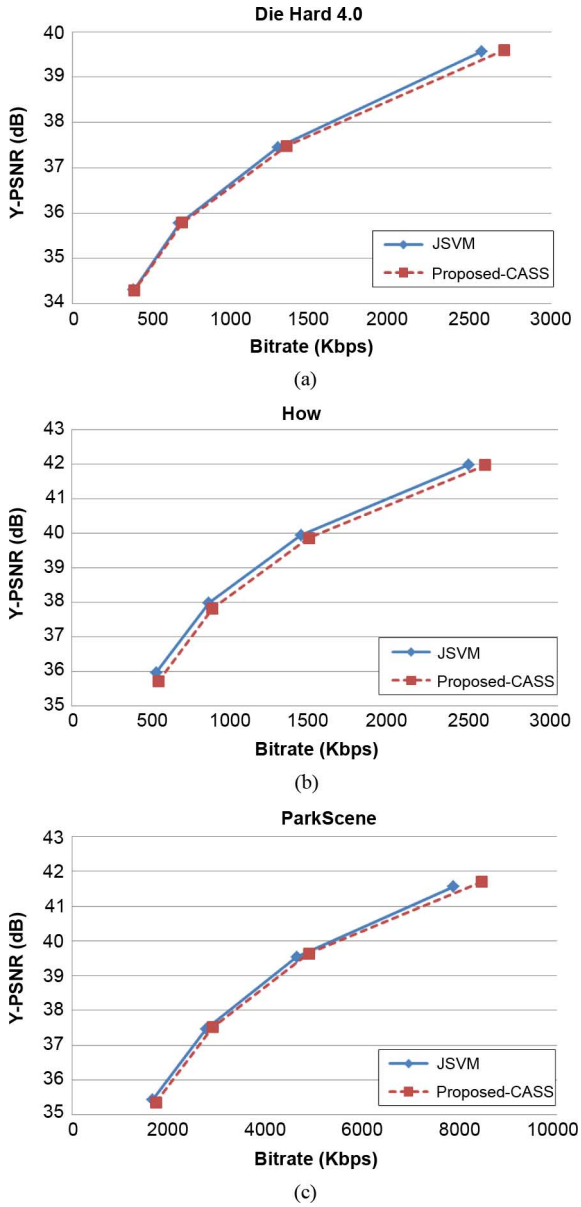

Fig. 9. Rate-distortion performance comparisons between the proposed framework and the conventional SVC for the high-resolution layer video.The GOP size is set to 16, and the QP is set to $(QP^1_{BL}, QP^1_{EL} = \{(32,36),(28,32),(24,28),(20,24)\}$ for (a) Diehard 4.0; (b) How; (c) Parkscene.

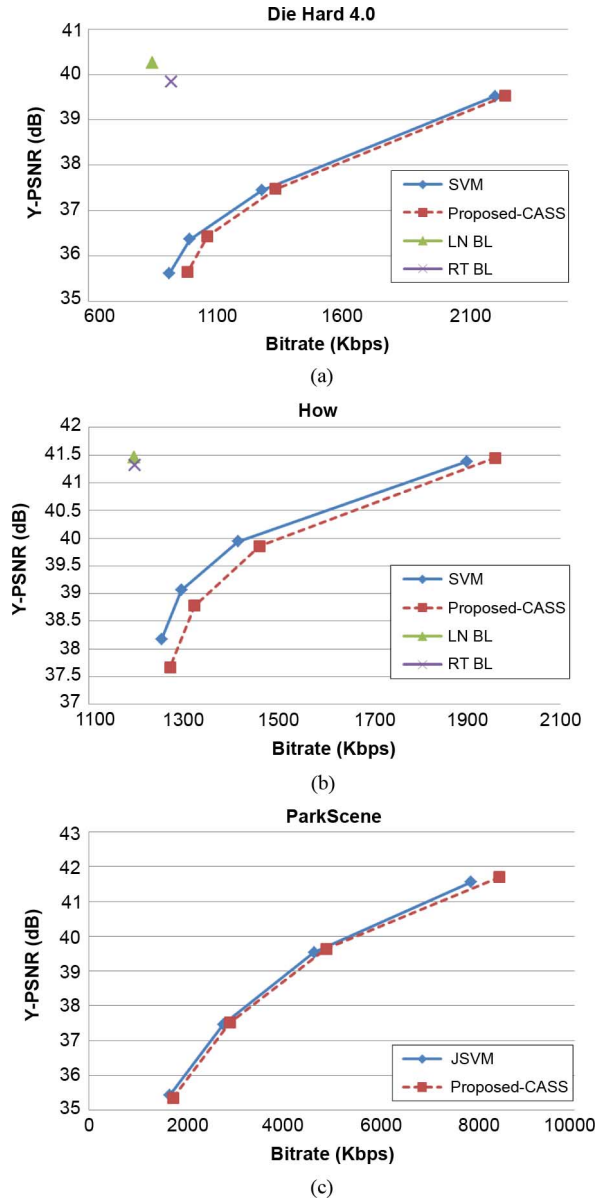

Fig. 10. Rate-distortion performance comparisons between the proposed framework and the conventional SVC for the high-resolution layer video.The GOP size is set to 16, and the QP is set to $(QP^2_{BL}, QP^2_{EL}) = \{(24,36),(24,32),(24,28),(24,24)\}$ for (a) Diehard 4.0; (b) How; (c) Parkscene.

with two spatial resolutions with a GOP size of 16. Each HR sequence is downscaled by three spatial downscaling schemes to obtain its LR sequences (i.e., the base-layer sequences), including uniform downscaling for SVC, the retargeting method proposed in [8] which was used in [15], and our retargeting method for the proposed CASS-SVC scheme.

To verify the performance of the proposed non-homogeneous inter-layer predictive coder under different quantization parameter (QP) settings, we designed two set of QP settings: $(QP^1_{BL}, QP^1_{EL}) = \{(32,36),(28,32),(24,28),(20,24)\}$ and $(QP^2_{EL}) = \{(24,36),(24,32),(24,28),(24,24)\}$. Note, in the former QP set, the QP value for the LR layer decreases as that for the HR layer decreases, whereas, in the latter QP set, the QP value for the base layer is kept the same regardless of the QP value for the HR layer. Figs. 9 and 10 compare the rate-distortion performances of the HR-layer video between our method and the conventional SVC[1] using the two QP settings, respectively. The blue solid lines indicate the R-D curves of the SVC, and the red dash lines indicate the R-D curves of the proposed method. Based on the proposed EL-BL mapping matrix, the non-homogeneous inter-layer prediction tools provide good prediction quality. The results in Fig. 9 show that, no matter how the change of the QP value in the base layer, our method only leads to slight quality degradation, which is due to the additional overhead required for sending the side information to the decoder. As shown in Fig. 10, for the second set of QP settings our method leads to average quality loss of 0.5 dB for low bit rate environment. The quality loss in the low bit-rate condition is mainly caused by the overhead of the side information, but while increasing the coding bit-rate, the

TABLE IV
COMPARISON OF OVERHEAD COSTS OF OUR METHOD AND THE METHOD PROPOSED IN [17] UNDER THE SAME QUALITY CONDITIONS

Sequence	PSNR of EL in dB	The method in [15]		The proposed CASS-SVC	
		Aggregate bitrate w/o side information in kbps	Aggregate bitrate in kbps	Aggregate bitrate w/o side information in kbps	Aggregate bitrate in kbps
ParkScene	37.5 dB	2480.4 (5.2%)	2536.6 (7.2%)	2406.4 (2.1%)	2406.7 (2.1%)
Kimono	38.7 dB	2123.1 (9.3%)	2198.3 (13.2%)	2080.3 (7.1%)	2082.7 (7.2%)

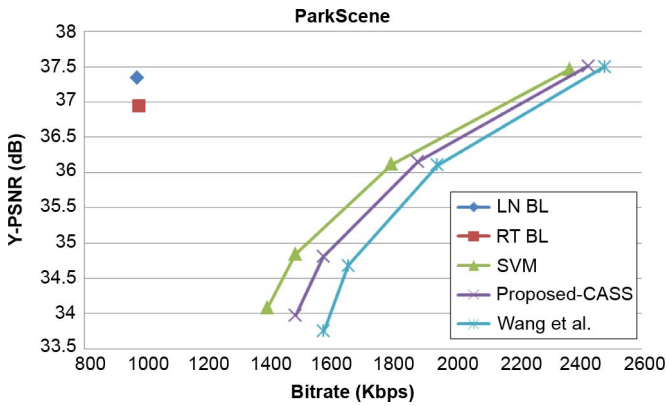

Fig. 11. Rate-distortion performance comparison of the EL video layer between the SVC, the proposed method and the method in [17]. The GOP size is set to 16, and the QP is set to $(QP^3_{BL}, QP^3_{EL}) = \{(30,42),(30,38),(30,34),(30,30)\}$.

effect of the overhead becomes insignificant. Table III shows the coding performance measured by BDPSNR (Bjontegaard Delta PSNR) and BDBR (Bjontegaard Delta Bit Rate) [27]. The average BDBR increments of the proposed method in the two QP sets are only 4.91% and 3.92%. In general, compared to the conventional SVC, thanks to the content-aware retargeting, the proposed method preserves significantly more important visual information in the LR-layer video, while maintaining comparable visual quality of the HR-layer video. Note, although the PSNR qualities of HR videos reconstructed by the proposed CASS-SVC is slightly lower than that reconstructed by the conventional SVC, our evaluation results [28] show that the proposed method achieves better subjective quality compared to SVC. This is because, similar to ROI-based coding, CASS-SVC better preserves the salient regions in the LR video, thereby achieving better visual quality perceptually after inter-layer prediction.

We also compare the coding performance of the proposed coder with that of the coder in[15] in the application scenario of supporting spatial scalability for both high definition (HD) TV and standard definition (SD) TV, where two 1280×720 HD sequences *ParkScene* and *Kimono* with an aspect ratio of 16:9 are downscaled to 720×576 SD sequences with an aspect ratio of 4:3. The GOP size is set to 16 and the used QP set is $(QP^3_{BL}, QP^3_{EL}) = \{(30,42),(30,38),(30,34),(30,30)\}$, the same setting as in [15]. Fig. 11 compares the R-D performances of SVC, our CASS-SVC, and the method in[15], evidently showing that our method outperforms the method in[15] in coding efficiency at all bitrates, especially at low bitrates where the overhead cost becomes relatively significant. The main reason is our method directly derives the frame-level scaling maps in a shot from the shot-level scaling map instead of sending per-frame scaling maps as in [15], leading to much fewer side information. As shown in Table IV, under the same PSNR quality condition, thanks to the shot-based scheme and the efficient interlayer prediction, our method achieves effective bitrate saving by about 5.1% and 6.0% of the overall bit-rate, compared to the method in[15]. The additional bitrates for side information with our method are only about 0.3 kbps for *ParkScene* and 2.4 kbps for *Kimono*, whereas the compared method consumes 56.2 kbps and 75.2 kbps for the two videos, respectively, which justifies the drastic overhead reduction with our scheme. For subjective evaluation, one can obtain the HR and LR video clips from our project website [28].

VII. CONCLUSION

We proposed a novel content-adaptive spatial scalability coding framework for SVC. The proposed framework consists of three modules to preserve the important content in the retargeted LR video without sacrificing the coding efficiency for the HR layer. We also proposed a new shot-based video retargeting method to successfully achieve important information preservation while maintaining good spatio-temporal coherence at a low overhead cost. Based on our framework, we have proposed a side information coder and efficient non-homogeneous interlayer prediction coding tools to achieve good coding efficiency in the HR layer. Thanks to the shot-based retargeting approach, our results show that, compared to existing schemes, while maintaining comparable visual quality for the LR video, the proposed method consumes much lower overhead bitrate, thereby achieving better rate-distortion performance in coding the HR video.

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable extension of the H.264/MPEG-4 AVC video coding standard," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.
[2] C. A. Segall and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 17, no. 9, pp. 1121–1135, Sep. 2007.
[3] A. Shamir and O. Sorkine, "Visual media retargeting," *ACM SIGGRAPH ASIA Courses (SIGGRAPH ASIA '09)*, pp. 1–13, 2009.
[4] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, p. 16, 2008.
[5] D. Han, X. Wu, and M. Sonka, "Optimal multiple surfaces searching for video/image resizing—a graph-theoretic approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, 2009, pp. 1026–1033.
[6] S. Kopf, J. Kiess, H. Lemelson, and W. Effelsberg, "FSCAV-fast seam carving for size adaptation of videos," in *Proc. ACM Int. Conf. Multimedia*, Beijing, China, Oct. 2009, pp. 321–330.
[7] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, 2007, pp. 1–6.
[8] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph.*, vol. 28, no. 5, 2009.
[9] Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee, "Scalable and coherent video resizing with per-frame optimization," *ACM Trans. Graph.*, vol. 30, no. 4, 2011.

[10] T.-C. Yen, C.-M. Tsai, and C.-W. Lin, "Maintaining temporal coherence in video retargeting using mosaic-guided scaling," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2339–2351, Aug. 2011.
[11] C.-M. Tsai, T.-C. Yen, and C.-W. Lin, "Mosaic-guided video retargeting for video adaptation," in *Proc. Conf. Applicat. Digital Image Process. XXXIV, SPIE Optics + Photonics 2011*, San Diego, CA, USA, Aug. 2011.
[12] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187–198, Feb. 2012.
[13] M. Décombas, F. Capman, E. Renan, F. Dufaux, and B. Pesquet-Popescu, "Seam carving for semantic video coding," in *Proc. Conf. Appl. Digital Image Process. XXXIV, SPIE Optics + Photonics 2011*, San Diego, CA, USA, Aug. 2011.
[14] M. Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, and F. Capman, "Improved seam carving for semantic video coding," in *Proc. IEEE Workshop Multimedia Signal Process*, Banff, AB, Canada, Sep. 2012, pp. 53–58.
[15] Q. Wang, R. Hu, and Z. Wang, "Intracoding and refresh with compression-oriented video epitomic priors," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 22, no. 5, pp. 714–726, May 2012.
[16] Y. Wang, N. Stefanoski, M. Lang, A. Hornung, A. Smolic, and M. Gross, "Extending SVC by content-adaptive spatial scalability," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 3493–3496.
[17] A. Smolic, Y. Wang, N. Stefanoski, M. Lang, A. Hornung, and M. H. Gross, "Non-linear warping and warp coding for content-adaptive prediction in advanced video coding applications," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, China, Sep. 2010, pp. 4225–4228.
[18] L. Shen and Z. Zhang, "Content-adaptive motion estimation algorithm for coarse-grain SVC," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2582–2591, May 2012.
[19] C.-W. Su, H.-Y. M. Liao, H.-R. Tyan, C.-W. Lin, D.-Y. Chen, and K.-C. Fan, "Motion flow-based video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1193–1201, Oct. 2007.
[20] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, Sep. 2004.
[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[22] R. Szeliski, "Image alignment and stitching: a tutorial," *Found. and Trends in Comput. Graph. Vis. (FTCGV)*, vol. 2, no. 1, pp. 1–104, 2006.
[23] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *ACM Commun.*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[24] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optimiz.*, vol. 9, no. 4, pp. 877–900, 1999.
[25] S. Mehrotra, "On the implementation of a primal-dual interior point method," *SIAM J. Optimiz.*, vol. 2, pp. 575–601, 1992.
[26] A. Segall, "Upsampling and down-sampling for spatial scalability," *Joint Video Team, Doc. JVT-R070*, Jan. 2006.
[27] G. Bjontegaard, "Calculation of average PSNR difference between RD curves," *ITU-T Q.6/16, Doc. VCEG-M33*, Apr. 2001.
[28] NTHU Video Scaling project, [Online]. Available: http://www.ee.nthu.edu.tw/cwlin/cass_svc/cass_svc.htm

Chia-Wen Lin (S'94–M'00–SM'04) received his Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000.

He is currently an Associate Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, during 2000–2007. Prior to joining academia, he worked for the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, during 1992–2000. His research interests include video content analysis and video networking.

Dr. Lin is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the *IEEE Multimedia*, and the *Journal of Visual Communication and Image Representation*. He is also an Area Editor of *EURASIP Signal Processing: Image Communication*. He has been Chair of Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society since September 2013. He served as Technical Program Co-Chair of the IEEE International Conference on Multimedia & Expo (ICME) in 2010, and Special Session Co-Chair of the IEEE ICME in 2009. He was a recipient of the 2001 Ph.D. Thesis Awards presented by the Ministry of Education, Taiwan. His paper won the Young Investigator Award presented by VCIP 2005. He received the Young Faculty Awards presented by CCU in 2005 and the Young Investigator Awards presented by National Science Council, Taiwan, in 2006.

Chia-Ming Tsai received the B.S. degree from Feng Chia University, Taichung, Taiwan, in 2003, and the M.S. and Ph.D. degrees from National Chung-Cheng University, Chiayi, Taiwan, in 2005 and 2013, respectively, all in computer science and information engineering.

Dr. Tsai served as a software engineer with CyberLink Inc., Taipei, Taiwan, from August 2012 to March 2013. He has worked for InterDigital Inc., San Diego, USA, as an Intern since May 2013. His research interests include video coding and video content adaptation.

Po-Chun Chen received the B.S. degree from National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan, in 2010, and the master degree from National Tsing Hua University, Hsinchu, Taiwan, in 2012, both in Electrical Engineering.

He joined the Display Technology Development Department of Faraday Technology Corporation as a senior engineer in October 2012. His research interests include video coding and image/video signal processing.