Maintaining Temporal Coherence in Video Retargeting Using Mosaic-Guided Scaling

Tzu-Chieh Yen, Chia-Ming Tsai, and Chia-Wen Lin, *Senior Member, IEEE*

***Abstract*—Video retargeting from a full-resolution video to a lower resolution display will inevitably cause information loss. Content-aware video retargeting techniques have been studied to avoid critical visual information loss while resizing a video. Maintaining the spatio-temporal coherence of a retargeted video is very critical on visual quality. Camera motions and object motions, however, usually make it difficult to maintain temporal coherence using existing schemes. In this paper, we propose the use of a panoramic mosaic to guide the scaling of corresponding regions of video frames in a video shot to ensure good temporal coherence. In the proposed method, after aligning video frames in a shot to a panoramic mosaic constructed for the shot, a global scaling map for these frames is derived from the panoramic mosaic. Subsequently, the local scaling maps of individual frames are derived from the global map and is further refined according to spatial coherence constraints. Our experimental results show that the proposed method can effectively maintain temporal coherence so as to achieve good visual quality even a video contains camera motions and object motions.**

***Index Terms*—Spatio-temporal coherence, video adaptation, video retargeting, video scaling.**

I. Introduction

With the rapid growth of handheld devices and wireless networks, sharing media content through these devices becomes more and more popular. The display size of a handheld device is typically much smaller than that of a TV or of a computer monitor. Spatial video scaling is therefore required to adapt visual content for the display formats of these handheld devices. However, uniform downsizing usually makes major objects too small to be recognized well. Moreover, the aspect ratio of a film is usually different from that of the display of a TV or a handheld device, making it necessary to scale or crop a video to adjust the aspect ratio. Fig. 1(b)–(d) shows three typical video resizing methods, the letterboxing, uniform scaling, and cropping methods, that are widely used in video processing applications. No matter how the visual content is resized to another lower resolution, it cannot prevent information loss from its full-resolution version.

Manuscript received May 24, 2010; revised September 28, 2010, January 28, 2011; accepted February 01, 2011. Date of publication February 14, 2011; date of current version July 15, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bulent Sankur.

T.-C. Yen is with ASUSTek Computer Inc., Taipei 11259, Taiwan.

C.-M. Tsai is with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 62102, Taiwan.

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

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/TIP.2011.2114357

Video retargeting is a structure-level video adaptation technique that resizes a video from one resolution to another lower resolution without severely deforming major content. An ideal video retargeting method has to preserve major visual content and avoid critical visual information loss while resizing the visual content [1]. To address this problem, several content-aware video retargeting methods have been proposed. According to the granularity of processing unit, these methods can be classified into three kinds of approaches: pixel-based approaches [2]–[5], region/patch-based approaches [6]–[10], and object-based approaches [11], [12]. We shall introduce these methods in more detail in Section II.

Although several content-aware image retargeting methods [2], [13]–[15] have proven to achieve good visual quality in resizing a single image, directly extending these image-based retargeting methods to video applications usually causes severe temporal incoherence artifacts. This is because the image-based retargeting schemes deal with the resizing of video frames separately without taking into account the temporal correlation of neighboring frames, leading to variation of the scaling factor of a corresponding region in neighboring frames which causes visually annoying artifacts on the region such as stretching (the reverse of stretching), shrinking (repeated stretching and shrinking), and waving (repeated stretching and shrinking). Although several video retargeting methods have been proposed to address the temporal incoherence problem, camera motions and object motions make it difficult to maintain temporal coherence with existing video retargeting schemes. With camera motions, a region would move to different spatial locations of neighboring frames. If a video retargeting method does not properly consider the spatio-temporal relationship, the scaling factor for the region may vary significantly in neighboring frames. Besides, a significant object movement or deformation in neighboring frames will puzzle video retargeting as well. For example, in the case that a video object moves from right to left in still background, initially, when the object stays at the right side, a content-aware retargeting method tends to trim the left side background at the first few video frames, but when the object moves to the left side, the resizing operator turns to trim the right side background. The inconsistent reduction of the background regions causes the jittery artifact and stretching/shrinking artifact in the video.

Video retargeting has several possible application scenarios. It can be performed at the decoder side, at the encoder side, or at both sides in an online or offline manner. Different scenarios will impose different constraints on video retargeting. For example, if the aim is to achieve online retargeting at the receiver side, the real-time processing requirement would con-

Fig. 1. Example of resizing (a) an original video frame from 16:9 to 4:3 by using (b) the letterboxing method, (c) the uniform scaling method, (d) the cropping method, and (e) the proposed method.

strain the complexity of the retargeting scheme, thereby influencing the output video quality. If the intended application is based on the assumption of offline retargeting the content at the encoder side for specific devices, then a more sophisticated scheme can be used to provide better visual quality, compared to the online approaches. However, the offline nature would reduce the flexibility/adaptivity of the output resolution. For retargeting a prestored video, one hybrid approach is to first perform some offline processing (e.g., feature extraction, saliency detection, frame registration and mosaicking) on the video at the encoder side, and then store the resulting output as metadata. The metadata are subsequently used to significantly reduce computation while performing online retargeting at the encoder/decoder, thereby relaxing the complexity constraint as well as achieving a good tradeoff between visual quality, format flexibility, and online complexity. Our method is aimed at enhancing the quality of retargeted video for applications that allow offline processing at the encoder side.

Our primary goal is to solve the temporal incoherence problem in a systematic way, rather than resorting to numerous temporal coherence constraints which are usually content dependent and are difficult to draw a unified set of constraints which are suitable for all types of videos. To ensure good temporal coherence, our proposed method first constructs a panoramic mosaic for a video shot and then uses it to derive a global scaling map for the frames of the shot. This global scaling map then serves as a guideline for deriving local scaling maps of individual video frames in the shot so as to ensure in the resized video the coherence of the scaling factors for each corresponding region in neighboring frames. The local scaling maps are further refined subject to a set of spatial coherence constraints to avoid spatial geometric distortions.

The rest of this paper is organized as follows. Section II summarizes the state-of-the-art content-aware video retargeting approaches. Section III gives a general formulation of non-homogeneous video resizing based on a spatio-temporal optimization framework. Our proposed mosaic-guided scaling method is presented in Section IV. Section V reports and discusses the experimental results. Finally, conclusions are drawn in Section VI.

II. Related Work

Several content-aware video retargeting methods have been proposed in recent years. These methods mainly aim to retain as much human interested regions as possible in a spatially downscaled video by trimming unimportant content, thereby preserving in the resized video the main concept inside the source video. The video retargeting methods can be classified into three kinds, namely, pixel-based approaches, region/patch-based approaches, and object-based approaches. Generally, a content-aware video retargeting method consists of two parts: energy function and resizing algorithm. The energy function which, in most existing works, is constituted of low-level perceptual features (e.g., gradient, color, and motion) to discover visually important regions of a video frame. Accrodingly, the resizing algorithm trims video frames non-homogeneously based on the energy values of pixels, patches, regions, or objects.

The pixel-based approaches resize video frames in the pixel domain. The seam-carving-based methods are among the most representative pixel-based approaches [2], [3]. Based on an energy function, the methods continuously remove a spatio-temporal surface until reaching the desired video resolution. Several variants of seam carving have been proposed to improve the visual quality by finding suitable low-energy spatio-temporal cubes to discard, or to reduce computational complexity [16]–[18]. However, with complex camera and object motions, finding a surface that does not disturb important video content becomes difficult.

Several warping-based video retargeting schemes [4], [5], [19] also belong to the pixel-based class. Wolf *et al.* [4] 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, whereas the distances of an important pixel to its neighboring pixels is retained. However, this method is only optimized at a desired resolution. It needs to recompute the shrinkability of each pixel when imposing another resolution constraint, making it impractical for real-time applications that require resolution change. To address this problem, Zhang *et al.* [19] improved the method by defining a per-pixel cumulative shrinkability map to scale each frame. The shrinkability map describes how close a pixel can approach to its neighboring pixels. In the method, it is not necessary to perform full computation when resizing a video to another video resolution, thereby achieving computation saving. To improve temporal coherence, Krähenbühl *et al.* [5] proposed to take into account the influence of scene change and object motion in a video. The method first uses a scene cut detector to detect discontinuities in the video and then computes bilateral temporal coherence energy accordingly for warp computation. Besides, it uses temporal filtering of per-frame saliency maps over a time window to account for the future changes of salient regions.

The region/patch-based approaches divide each video frame into many regions/patches. The scaling factor (or sampling rate)

of each region/patch is determined by a spatio-temporal optimization process. Kim *et al.* [6] proposed to split an image into many strips. The optimal scale of each strip is then determined based on the Fourier analysis. In this method, a video sequence is treated as a spatio-temporal cube. The cube is subsequently divided into many individual regions and the corresponding sampling rate for each region is determined according to the region's importance. In [7], Shi *et al.* proposed a context-assisted video retargeting scheme that combines the high-level visual concepts and visual attention into a spatio-temporal importance map. The importance map is then incorporated with their proposed 3-D rectilinear grid resizing scheme. The performance of the method was evaluated on sports and advertisement videos. The cropping-based methods proposed in [8] and [9] define a target region that includes the most important part of the original video. The target region must have the same size of the expected resolution. The cropping-based method also needs to maintain the temporal coherence of the cropped regions to prevent the jittery artifact. The main weakness of cropping-based method is that the discarded regions often still contain important information.

The object-based approaches segment a video frame into foreground objects and background [11], [12]. The objects and background are then resized by different resizing techniques. The object-based schemes rely on accurate object segmentation to extract all possible objects. With the foreground and background masks, individual objects are recomposed to the desired video sizes. However, inaccurate object segmentation will cause perceptually unpleasant artifacts along the boundary of an object.

A few video retargeting methods use image registration techniques to mitigate the negative impact of object and camera motions on temporal coherence [18], [10]. Image registration aligns video frames by fitting a camera motion model between consecutive frames. The geometrical correspondence between every two consecutive frames is then established based on the estimated camera motion. Kopf *et al.* [18] proposed to construct a panoramic mosaic to track the (local) object motions and (global) camera motions. Based on the concept of spatio-temporal cube, the panoramic mosaic is used to identify robust seams to remove so as to preserve temporal coherence. However, when the object movement covers a large portion of a frame, only few robust seams can be found for video resizing. Wang *et al.* [10] proposed a method of achieving motion-aware temporal coherence for video retargeting. The method also uses frame alignment to tackle the problem of camera and object motions. In order to track important content across neighboring frames, frame alignment is performed to blend the importance (saliency) map. The estimated camera motions are subsequently used to constrain the object and camera motions as well as to prevent content deformation. However, it may produce false camera motion due to an insufficient number of frames used to blend the importance map.

Different from the existing schemes, our proposed mosiac-guided scaling scheme is a hybrid approach. Our scheme first constructs a panoramic mosaic from a spatio-temporal cube (e.g., a video shot) to record the object and camera motions. The panoramic mosaic is then used to derive the shot-level global scaling map. The local scaling map of each frame is first extracted from the global scaling map after aligning the frame to the mosaic, and is further refined subject to predefined spatial coherence constraints. Each frame is resized according to its local scaling map. The proposed mosaic-guided retargeting approach gracefully maintains temporal coherence by making global decision of scaling factors so as to mitigate the influence of object and camera motions.

TABLE I
NOTATION

Symbols	Meanings
W, H	The width and height of the original video
W', H'	The width and height of the resized video
t	The time index of current frame
n	The index of iterative optimization
$\mathbf{m} = (i,j)$	The coordinate of a pixel/patch in the original frame
$\mathbf{m}' = (i',j')$	The corresponding coordinate of $\mathbf{m}$ in the resized frame
D_S, D_T, and D_{ST}	The spatial, temporal, and spatio-temporal incoherence distortions
D_{info}	The information loss of a video frame after resizing
$e^{(t)}(i,j)$	The energy value at the (i,j)-th pixel of the t-th frame
$\mathbf{H}^{(t)}$	The projective transform for the t-th frame
D_{GM}	The temporal coherence constraint
D_{ROI}	The region-of-interest (ROI) deformation constraint
D_{SS}	The spatial smoothness constraint
$\mathbf{s}_s\left((i',j')_M\right)$	The set of scaling factors for pixel (i',j') of the panoramic mosaic
$s_G\left((i',j')_M\right)$	The scaling factor for pixel (i',j') of the global scaling map
$s_L^{GM}\left((i,j)_{\text{in}}^{(t)}\right)$	The local scaling maps derived from the global scaling map
$s_L^{ini}(i,j)$	The initial local scaling factor of pixel (i,j) of a video frame
$s_L^{(n)}\left((i,j)_{\text{in}}^{(t)}\right)$	The local scaling map obtained from the n-th round iteration
s_L^*	The final refined scaling maps of individual frames

III. Formulation of Non-Homogeneous Video Resizing

The symbols used in this paper are listed in Table I. Assume we resize a video from resolution to $W \times H$ to $W' \times H'$, where W and H are the width and height of the original video, and W' and H' are the width and height of the resized video. Suppose that a spatio-temporal cube (e.g., a video shot) consists of N frames which are denoted as $\mathbf{I}_{\text{in}} = \{I_{\text{in}}^{(t)}\}_{t=1}^{N}$ and the corresponding resized frames are denoted as $\mathbf{I}_{\text{out}} = \{I_{\text{out}}^{(t)}\}_{t=1}^{N}$. Video retargeting is to find a transform $I_{\text{out}}^{(t)} = \mathbf{T}(I_{\text{in}}^{(t)})$ which can preserve in the resized frame the most important content while maintaining spatio-temporal coherence.

Generally, the video retargeting can be formulated as an optimization problem by which the optimal retargeted video $\mathbf{I}_{\text{out}}^*$ can be obtained as

$$\mathbf{I}_{\text{out}}^* = \arg\min_{\mathbf{I}_{\text{out}}} \left\{ \sum_{t=1}^{N} \left(D_{ST}\left(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}, I_{\text{out}}^{(t-1)}\right) + \lambda D_{\text{info}}\left(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}\right)\right) \right\} \quad (1)$$

where $D_{ST}(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}, I_{\text{out}}^{(t-1)})$ denotes the spatio-temporal distortion, $D_{\text{info}}(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)})$ denotes the information loss, and λ is a weighting factor.

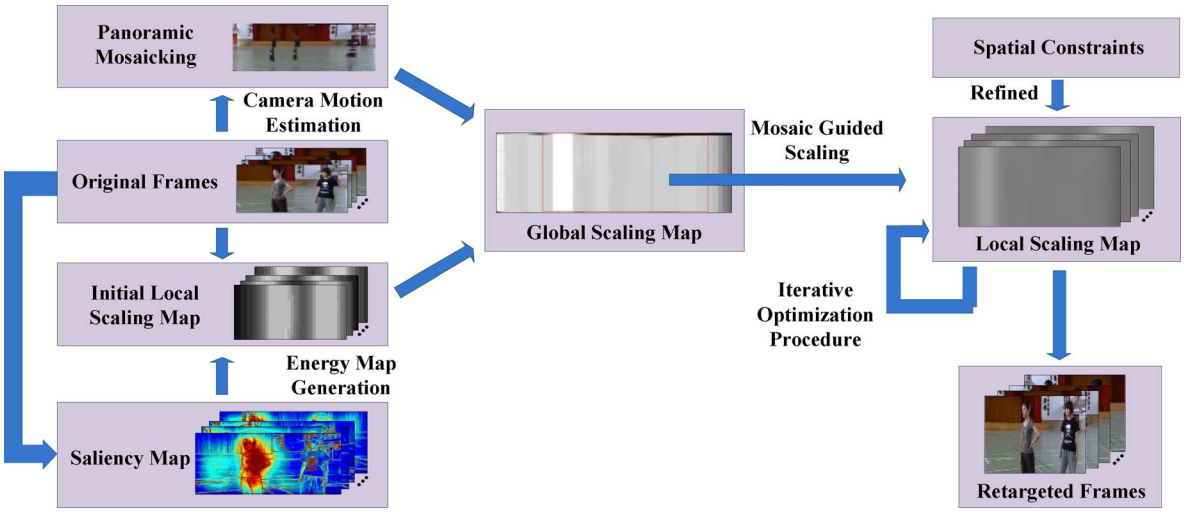

Fig. 2. Flow diagram of the proposed method.

The information loss after resizing a frame can be measured by the energy distortion between the original frame and the resized one as follows:

$$D_{\text{info}}\left(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}\right) = E\left(I_{\text{in}}^{(t)}\right) - E\left(I_{\text{out}}^{(t)}\right) \tag{2}$$

where $E(I_{\text{in}}) = \sum_{j=1}^{H}\sum_{i=1}^{W} e(i,j)$ and $E(I_{\text{out}}) = \sum_{j=1}^{H}\sum_{i=1}^{W} e(i,j)\cdot s_L(i,j)$ denote the total energy values of the original and resized frame, respectively. $e(i,j)$ and $s_L(i,j)(0 \le s_L(i,j) \le 1)$, respectively, represent the energy value and the scaling factor of pixel (i,j) in a video frame.

The spatio-temporal distortion $D_{ST}(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}, I_{\text{out}}^{(t-1)})$ can be further divided into two terms:

$$D_{ST}\left(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}, I_{\text{out}}^{(t-1)}\right) = D_S\left(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)}\right) - D_T\left(I_{\text{out}}^{(t-1)}, I_{\text{out}}^{(t)}\right) \tag{3}$$

where $D_S(I_{\text{in}}^{(t)}, I_{\text{out}}^{(t)})$ and $D_T(I_{\text{out}}^{(t-1)}, I_{\text{out}}^{(t)})$ denote the spatial and temporal incoherence distortions, respectively.

Let the spatial structural value of a pixel be defined as the sum of the distances between the pixel and its neighborhood (e.g., the right, left, upper, and lower pixels). The spatial incoherence distortion [4] can be measured by the spatial structural deformation defined as the sum of the difference of the spatial structural values of individual pixels in the original frame ($SS_{\text{in}}(\mathbf{m})$) and of their corresponding pixels in the resized frame ($SS_{\text{out}}(\mathbf{m}')$), weighted by the pixels' energy values as follows:

$$D_S(I_{\text{in}}, I_{\text{out}}) = \sum_{\mathbf{m}} \left\{ \left(\underbrace{\sum_{N(\mathbf{m})} \|\mathbf{m} - N(\mathbf{m})\|}_{SS_{\text{in}}(\mathbf{m})} - \underbrace{\sum_{N(\mathbf{m}')} \|\mathbf{m}' - N(\mathbf{m}')\|}_{SS_{\text{out}}(\mathbf{m}')} \right) \cdot e(\mathbf{m}) \right\} \tag{4}$$

where $\mathbf{m} = (i,j)$ denotes the coordinate of a pixel/patch of the original frame, $e(\boldsymbol{\gamma})$ is the pixel's corresponding energy value, which is nonnegative (will be explained in (6)), $\mathbf{m}' = (i',j')$ denotes the corresponding coordinate of $\mathbf{m}$ in the resized frame, and $N(\mathbf{m})$ and $N(\mathbf{m}')$ represent the neighborhoods of $\mathbf{m}$ and $\mathbf{m}'$, respectively. The metric $\|\mathbf{m} - N(\mathbf{m})\|$ denotes the Euclidean distance between $\mathbf{m}$ and $N(\mathbf{m})$. For simplicity, the time index t is omitted in (4). Note, in video downscaling, $SS_{\text{in}}(\mathbf{m}) \ge SS_{\text{out}}(\mathbf{m}')$ and $e(\mathbf{m}) \ge 0$. Therefore, $D_S(\cdot)$ is nonnegative.

The temporal incoherence distortion can be measured by the geometrical distortion between two consecutive resized frames:

$$D_T\left(I_{\text{out}}^{(t-1)}, I_{\text{out}}^{(t)}\right) = I_{\text{out}}^{(t-1)} - \delta_{t\mapsto t-1} I_{\text{out}}^{(t)} \tag{5}$$

where $\delta_{t\mapsto t-1}$ denotes the geometrical coordinate mapping from the tth frame to the $(t-1)$th frame.

IV. Proposed Video Retargeting Scheme

As illustrated in Fig. 2, the proposed mosaic-guided scaling scheme consists of five major operations: the proposed mosaic-guided scaling scheme consists of five major operations: energy map generation, shot-level panoramic mosaic construction, global scaling map generation, local scaling map generation, and frame resizing. In the energy map generation unit, the frame-level energy map, that indicates the visual importance of individual pixels, is obtained by using the energy function derived from the perceptual-quality significance map (PQSM) [20] and pixel gradients. Based on the energy map, a frame's initial local scaling map is obtained by maximizing the energy preserved in a resized frame subject to a set of spatial constraints. We use a linear programming solver to solve the constrained optimization problem to obtain the initial local scaling map. The shot-level panoramic mosaicking unit performs camera motion estimation and frame alignment to build a panoramic mosaic to record the geometrical structure in a video shot. Based on the panoramic mosaic and the initial local scaling maps of the shot, a global scaling map is derived to provide a global reference for

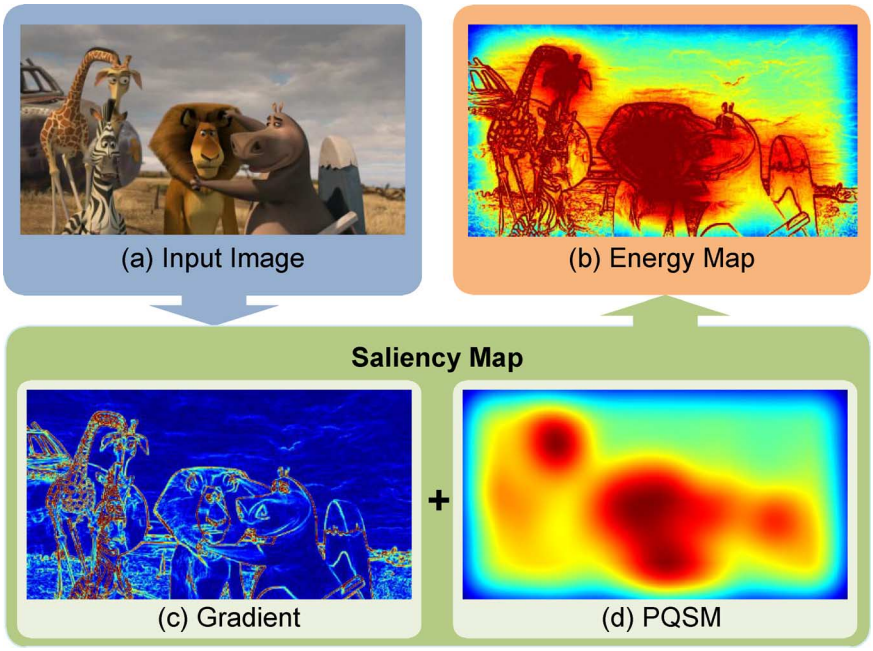

Fig. 3. Flow diagram of generating the energy map, where a saliency map derived from gradient values and PQSM [20] is used as the energy map. A pixel in dark red in (b) to (d) indicates that human eyes are more sensitive to the pixels.

achieving temporal coherence in the video shot. The final local scaling maps of the shot's individual frames are first extracted from the global scaling map and are then refined by imposing predefined spatial coherence constraints in an iterative manner. Consequently, the video frames are resized according to their corresponding local scaling map. The detailed operations of the proposed retargeting scheme are elaborated below.

A. Initialization

The proposed mosaic-guided scaling method needs three kinds of maps for resizing a video shot: the frame-level energy maps, the initial local scaling maps, and the shot-level panoramic mosaic.

1) Frame-Level Energy Maps: The energy function, which is used to represent the visual importance (saliency) of a pixel in each video frame, plays an important role in content-aware image/video retargeting. With an appropriate energy function, one is able to apply optimization techniques to minimize the energy loss caused by the removal of image content. Salient regions can be detected based on top-down and/or bottom-up visual attention models [21]. A top-down visual attention model is a goal-driven model which is related to cognitive psychology. On the other hand, a bottom-up visual attention model is a stimuli-driven model based on low-level features (e.g., colors, gradients, motion). For example, in [3] the gradient energy is used to localize visually important regions to preserve the spatial structure of an image. However, the gradient energy cannot fully capture the eye-sensitive regions. Besides the gradient energy, the method proposed in [4] utilizes more visual attention features, e.g., facial feature and motion feature, to improve the detection accuracy. In [6], Kim *et al.* used frequency domain features to localize human interested regions. In [15], the image-level saliency detection method proposed in [21] was adopted.

As shown in Fig. 3, the proposed method adopts the PQSM model [20] to generate the saliency map. PQSM consists of three steps, including visual attention features integration, post-processing, and motion suppression, to generate a visual sensitivity map. The saliency map generated by PQSM provides fairly accurate locations, whereas the detected region boundaries are not sharp enough, leading to difficulty in preserving the content structure. Therefore, we propose using an energy fusion function to combine the gradient energy and the PQSM-based saliency map as

$$e(i,j) = \alpha_1 \cdot \text{Gradient}(i,j) + \alpha_2 \cdot \text{PQSM}(i \cdot j) \tag{6}$$

where $e(i,j)$ represents the energy value of the (i,j)th pixel. The values of $\text{Gradient}(i,j)$ and $\text{PQSM}(i,j)$ are both normalized to [0, 1] using the min-max normalization. The two weights, α_1 and α_2 are both set as 0.5. Therefore, the energy value ranges within [0, 1].

2) Initial Local Scaling Maps: In frame resizing, a frame needs a local scaling map to determine how to scale a pixel (or a patch) non-homogeneously. Our method uses an energy-based frame resizing approach. The local scaling map of a frame is obtained by solving a constrained energy-preserving optimization problem that is to maximize the energy retained in a resized frame while maintaining the spatial coherence in the frame by

$$\begin{aligned} s_L^{\text{ini}*} &= \underset{\{s_L^{\text{ini}}(i,j)\}}{\arg\max}\, E(I_{\text{out}}) \\ &= \underset{\{s_L^{\text{ini}}(i,j)\}}{\arg\max} \sum_{j=1}^{H}\sum_{i=1}^{W} e(i,j)\cdot s_L^{\text{ini}}(i,j) \\ \text{s.t.} &\sum_i s_L^{\text{ini}}(i,j) = W', \quad \text{and} \\ &\left|S_L^{\text{ini}}(i,j) - S_L^{\text{ini}}(i,j+1)\right| \le \text{TH}_s, \quad \forall j \end{aligned} \tag{7}$$

where $s_L^{\text{ini}*}$ represents the optimal initial local scaling map, $e(i,j)$ denotes the energy magnitude at the (i,j)th pixel of the saliency map, $s_L^{\text{ini}}(i,j)$ denotes the initial local scaling factor for pixel (i,j) in the tth frame, where $0 \leq s_L^{\text{ini}}(i,j) \leq 1$, and W' is the target width. The threshold for the imposed spatial constraint $TH_S = 0.06$ for all input videos.

As a result, the frame resizing method determines for each pixel an appropriate scaling factor under the spatial coherence constraints. Note that, using the frame-level scaling maps to resize individual frames cannot avoid temporal incoherence artifacts. We thus only use the scaling maps obtained by (7) as initial maps to derive a global shot-level scaling map, and then use the global scaling map to obtain the final local scaling maps that can ensure temporal coherence by an iterative optimization approach.

3) Shot-Level Panoramic Mosaic: Typically, a panoramic mosaic is generated by using three steps: feature points detection, camera motion estimation, and frame registration. Our method uses SIFT [22] to select feature points in each video frame, because SIFT is robust to scaling change (e.g., zoom-in and zoom-out manipulations). Camera motion estimation has been extensively studied and there exist several sophisticated models [23]. For the sake of simplicity, we use a simplified affine model with only scaling and translation parameters. Although it cannot characterize all possible camera motions, our experiments show that the simplified model achieves reasonably good accuracy in constructing a panoramic mosaic for a video shot.

Camera motion estimation and frame registration are essential steps of constructing a panoramic mosaic. We use RANSAC [24] to estimate camera motion between neighboring frames. Although RANSAC can prevent false model fitting from ill-featured correspondences, when most part of a frame is occupied with foreground regions, the chosen feature correspondence set is probably taken from the foreground regions, leading to frame misalignment and a polluted panoramic mosaic. To avoid the problem, we filter out those ill-featured correspondences of foreground regions by resorting to the saliency map. If the saliency value of a feature correspondence is larger than a predefined threshold that will be defined in (14), it is likely to be an object point and therefore should be removed from the RANSAC computation. In the frame registration in a shot, the panoramic mosaic is generated by using the estimated camera motions of the frames.

B. Mosaic-Guided Video Retargeting

As mentioned in (5), the scaling factor change between the resized frames $I_{\text{out}}^{(t-1)}$ and $I_{\text{out}}^{(t)}$ should be constrained by fitting the mapping model $\delta_{t\mapsto t-1}$. To this end, a shot-level panoramic mosaic is used to maintain the temporal coherence of video resizing under camera and object motions. A shot-level scaling map is derived from the panoramic mosaic. The local scaling maps of individual frames are extracted from the global scaling map, and are further refined by imposing a few predefined spatial constraints in an iterative manner. In this section, we first introduce the method of generating the global scaling map and then present the iterative process of refining the scaling factors of individual frames.

1) Global Scaling Map: Directly extending an image retargeting method to video retargeting usually leads to temporal incoherence artifacts, especially when a video contains camera motions or large object motions. Due to the camera or object motions, the correspondence (a patch or a pixel) in neighboring frames may have different spatial locations, sizes, and shapes, thereby being scaled differently. Such inconsistent scaling for a corresponding patch/pixel in neighboring frames results in temporal incoherence artifacts such as stretching, shrinking, and waving of object or background. To prevent such temporal artifacts, the scaling factors of the same visual content should be kept as consistent as possible in neighboring frames. Besides, the aspect ratios of a foreground object in neighboring frames should be kept consistent as well.

After constructing a panoramic mosaic for a video shot, a global scaling map is derived to synchronize the scaling of a corresponding pixel/patch in different frames of the shot. Let $\mathbf{H}^{(t)}$ denote the projective transform of the tth frame, $(i,j)_{\text{in}}^{(t)}$ the coordinate of the (i,j)th pixel in the tth original frame, and $(i',j')_M$ the projected coordinate of $(i,j)_{\text{in}}^{(t)}$ in the mosaic after frame alignment. Then, the projection of a coordinate is given by $(i',j')_M = \mathbf{H}^{(t)}(i,j)_{\text{in}}^{(t)}$. The global scaling map is simply obtained as the union of scaling factors after the transformation, as expressed by

$$\mathbf{s}_s\left((i',j')_M\right) = \bigcup_t \bigcup_{(i,j)_{\text{in}}^{(t)} \to (i',j')_M} s_L^{\text{ini}}\left(\mathbf{H}^{(t)}(i,j)_{\text{in}}^{(t)}\right) \quad (8)$$

where $s_L^{\text{ini}}(i,j)$ denotes the initial local scaling factor of pixel (i,j) of a video frame as obtained from (7), and $\mathbf{s}_s((i',j')_M)$ represents a set of scaling factors corresponding to pixel (i',j') of the panoramic mosaic, as the union operation in (8) is a many-to-one mapping, that is, $(i',j')_M$ may correspond to the scaling factors from different video frames and different pixels of the original video.

To obtain a single-valued mapping, we choose the maximum scaling factor in the set defined in (8) as the scaling factor for pixel $(i',j')_M$ of the global scaling factor map as follows:

$$s_G\left((i',j')_M\right) = \max\left\{\mathbf{s}_s\left((i',j')_M\right)\right\}. \quad (9)$$

2) Global Map Constraint: After deriving the global scaling map, the first-round local scaling maps are extracted from the global scaling map by

$$s_L^{GM}\left((i,j)_{\text{in}}^{(t)}\right) = s_G\left(\mathbf{H}^{(t)}(i,j)_{\text{in}}^{(t)}\right) = s_G\left((i',j')_M\right). \quad (10)$$

The mosaic-derived local scaling maps themselves are temporally coherent since a corresponding pixel/patch of neighboring frames has consistent scaling factors, but may not preserve spatial coherence well. We therefore propose an iterative optimization scheme which uses the mosaic-derived local scaling maps as a start-point to obtain the final local scaling

maps subject to spatial coherence constraints. Therefore, the first-round local scaling maps are set as

$$s_L^{(1)}\left((i,j)_{\text{in}}^{(t)}\right) = s_L^{GM}\left((i,j)_{\text{in}}^{(t)}\right). \tag{11}$$

To ensure the temporal coherence offered by the mosaic-derived local scaling maps, we use the first-round maps to constrain the iterated local scaling maps by introducing the following distortion cost:

$$D_{GM} = \sum_{i,j} d\left(s_L^{(n)}\left((i,j)_{\text{in}}^{(t)}\right), s_L^{GM}\left((i,j)_{\text{in}}^{(t)}\right)\right)^2 \tag{12}$$

where $s_L^{(n)}((i,j)_{\text{in}}^{(t)})$ denotes the nth round scaling factor of the (i,j)th pixel in the tth frame. The distance function $d(\cdot)$ in (12) is defined as

$$d(a,b) = \frac{\max(a,b)}{\max(\min(a,b),\varepsilon)}. \tag{13}$$

where ε is a small positive value to avoid the division-by-zero error when $\min(a,b) = 0$.

3) Spatial Coherence Constraints: In the optimization process, we impose the following constraints to prevent the spatial incoherence distortion.

a) Regions-of-Interest (ROI) Deformation Constraint: In order to maintain spatial coherence, the scaling factors of the pixels/patches of visually important foregrounds/backgrounds should be made consistent. To do so, we define a set ROI consisting of pixels/patches that belong to ROI, and Non-ROI consisting of the pixels/patches that are not ROI. To separate ROIs, we define a threshold, $TH_{\mathbf{ROI}}$, that is empirically set to be 0.6 for all sequences, for the classification

$$\begin{cases} \{i,j\} \in \mathbf{ROI}, & \text{if } e(i,j) \geq TH_{\mathbf{ROI}} \\ \{i,j\} \in \mathbf{Non\text{–}ROI}, & \text{otherwise} \end{cases} \tag{14}$$

where the energy value $e(i,j)$ of the (i,j)th pixel (or patch) is calculated by (6).

To maintain the consistency of foreground object size, the following spatial scaling inconsistency should be minimized:

$$D_{\text{ROI}} = \sum_{\mathbf{m}\in\mathbf{ROI}} d\left(s_L^{(n)}\left((\mathbf{m})_{\text{in}}^{(t)}\right), s_L^{(n)}\left((\mathbf{m}+\mathbf{1})_{\text{in}}^{(t)}\right)\right)^2 \tag{15}$$

where $\mathbf{m} = (i,j) \in \mathbf{ROI}$ indicates the pixel/patch belonging to ROI, $\mathbf{1} = (0,1)$ for horizontal resizing and $\mathbf{1} = (1,0)$ for vertical resizing. The distortion function $d(\cdot)$ is defined in (13).

b) Spatial Smoothness Constraint: If two vertically (or horizontally) adjacent pixels/patches are resized in different factors, the vertical (or horizontal) structures will be distorted. To avoid such spatial structural distortion, we need to constrain the difference between the scaling factors of two spatially adjacent pixels/patches. Assuming an image is downscaled in the horizontal dimension, we limit the sum of the differences between the scaling factors of every two vertically adjacent pixels/patches on a line not to exceed a threshold TH_{SS} as follows:

$$D_{SS} = \sum_j \sum_i \left| s_L^{(n)}(i,j)_{\text{in}}^{(t)} - s_L^{(n)}(i,j+1)_{\text{in}}^{(t)} \right|. \tag{16}$$

4) Iterative Optimization Procedure: After obtaining the first round local scaling map $s_L^{(1)}((i,j)_{\text{in}}^{(t)})$, an iterative optimization procedure is performed to find a converged solution $s_L^*((i,j)_{\text{in}}^{(t)})$ subject to three smoothness constraints: (12), (15), and (16). The final refined scaling maps of individual frames are derived iteratively from (17) using an iterative optimization solver:

$$\begin{aligned} s_L^* &= \arg\min_{s_L^{(n)}} D_{\text{total}} \\ &= \arg\min_{s_L^{(n)}} (D_{GM} + \lambda_1 D_{\text{ROI}} + \lambda_2 D_{SS}). \\ &\quad \text{s.t.} \left| s_L^{(n)}(i,j)_{\text{in}}^{(t)} - s_L^{(n)}(i,j+1)_{\text{in}}^{(t)} \right| \leq TH_{SS} \\ &\qquad\qquad \forall i,j \end{aligned} \tag{17}$$

where λ_1 and λ_2 are the weighting factor for D_{ROI} and D_{SS}, respectively. In our implementation, we set D_{GM}, D_{ROI}, and D_{SS} equally important (i.e., $\lambda_1 = \lambda_2 = 1$). The threshold for spatial constraint TH_{SS}, similar to the case in (7), is set to be 0.06.

To solve (17), we use the interior-point solver [25] that is designed for solving a large-scale optimization problem. The solver has proven to be capable of solving a wide range of problems, even when ill-conditioning and non-convexity is present. However, the solution might become trapped in a local minimum as it is a gradient-descent based approach. On the other hand, only full search can guarantee to reach a global minimum, whereas its complexity is much higher. To reduce computation, in our implementation, the distortions (12), (15), and (16) are evaluated in a patch rather than in a pixel. The iterative procedure is summarized in Table II.

5) Frame Resizing Based on Local Scaling Maps: After obtaining the final local scaling maps, the resized frame is generated by the pixel fusion based image downscaling proposed in [14]. The method is summarized below. First, after resizing, each pixel in the image is treated as a component whose width is scaled from unity (the original pixel) to a fractional number (i.e., the scaling factor), assuming the resizing is performed horizontally. The value of a resized pixel (i.e., a unit width of the joined pixels) is obtained by the linear combination of the values of the pixels that compose the unit width weighted by the widths of the component pixels. As shown in Fig. 4 [14], when mapping the (i',j)th pixel and the $(i'+1,j)$th pixel to the corresponding locations of the resized frame (as indicated by the gray dashed lines), the (i',j)th pixel value of the resized frame is fused from three regions. The first region contains the (m,j)th pixel value of the input frame and its length is C_1. The second region is contributed by the linear combination of $k-1$ pixels and its length is C_2. The third is from the $(m+k,j)$th pixel value of

TABLE II
ITERATIVE ALGORITHM FOR COMPUTING THE LOCAL SCALING MAPS

Initialization

E : Energy-based saliency map

s_L^{ini} : The initial local scaling maps

$\mathbf{H}$: The panoramic mosaic mapping

Construct the global scaling map s_G

$$\mathbf{s}_s\left((i',j')_M\right) = \bigcup_{(i,j)_{in}} \bigcup_{(i',j')_M} (i,j)_{\mathrm{in}}^{(t)}\Big),$$

$$s_G\left((i',j')_M\right) = \max\left\{\mathbf{s}_s\left((i',j')_M\right)\right\}$$

Iterate

D_{GM} : Global Map Constraint

$$D_{GM} = \sum_{i,j} d\left(s_L^{(n)}\left((i,j)_{\mathrm{in}}^{(t)}\right), s_L^{GM}\left((i,j)_{\mathrm{in}}^{(t)}\right)\right)^2$$

D_{FD} : ROI Deformation Constraint

$$D_{ROI} = \sum_{\mathbf{m}\in\mathbf{ROI}} d\left(s_L^{(n)}\left((\mathbf{m})_{\mathrm{in}}^{(t)}\right), s_L^{(n)}\left((\mathbf{m}+1)_{\mathrm{in}}^{(t)}\right)\right)^2$$

D_{SS} : Spatial Smoothness Constraint

$$D_{SS} = \sum_j \sum_i \left| s_L^{(n)}(i,j)_{\mathrm{in}}^{(t)} - s_L^{(n)}(i,j+1)_{\mathrm{in}}^{(t)} \right|$$

for all frames *t*

Derive the first-round local scaling map $s_L^{(1)}\left((i,j)_{\mathrm{in}}^{(t)}\right) = s_L^{GM}\left((i,j)_{\mathrm{in}}^{(t)}\right)$

for $n \le Max\_Iteration$

$$s_L^* = \arg\min_{s_L^{(n)}} D_{\mathrm{total}} = \arg\min_{s_L^{(n)}} \left(D_{GM} + \lambda_1 D_{ROI} + \lambda_2 D_{SS}\right).$$

$$\text{s.t. } \left| s_L^{(n)}(i,j)_{\mathrm{in}}^{(t)} - s_L^{(n)}(i,j+1)_{\mathrm{in}}^{(t)} \right| \le TH_{SS}, \quad \forall i,j$$

until convergence

Obtain the optimal local scaling map of the *t*-th frame $s_L^*\left((i,j)_{\mathrm{in}}^{(t)}\right)$

until all frames are processed

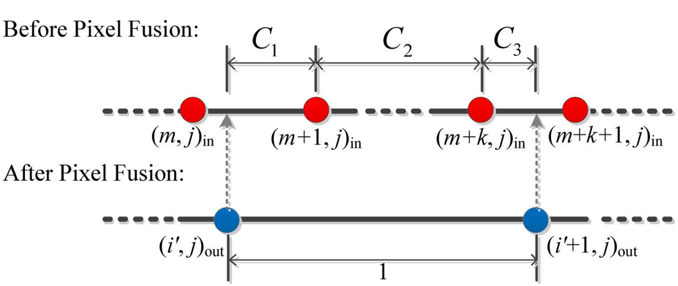

Fig. 4. Mapping of the (i', j)th pixel and the $(i'+1, j)$th pixel of the output frame to the corresponding locations of the resized frame.

the input frame and its length is C_3. The (i', j)th pixel value of the resized frame is computed by

$$I(i',j)_{\mathrm{out}} = (C_1 \cdot I(m,j)_{\mathrm{in}}) + \left(\sum_{l=1}^{k-1} s_L^*(m+l,j)_{\mathrm{in}} \cdot I(m+l,j)_{\mathrm{in}}\right) + (C_3 \cdot I(m+k,j)_{\mathrm{in}}) \quad (18)$$

In this paper, we only consider spatial resizing in one dimension. Two-dimensional scaling can be performed by two separable one-dimensional scaling operations. The process is to first obtain the shot-level panoramic mosaic, and then to use the mosaic to generate two global scaling maps: one for horizontal resizing and the other for vertical resizing based on the two sets of initial local scaling maps for horizontal and vertical downscaling, respectively. The two global maps are then used to generate two sets of refined local maps separately using the iterative optimization. As a result, the two-dimensional scaling can be done along one direction followed by the other. Note, there could be some spatial inconsistency of scaling for an object in the two directions, as the two directions are resized separately. Such inconsistency is not serious as long as the energy value of a foreground/background object does not vary significantly in the two directions.

V. EXPERIMENTS AND DISCUSSION

To evaluate the performance of our proposed method, we select test sequences that involve rich types of camera and object motions from cinema, drama, and animation videos. In the experimental settings, each test video is resized to the half size of the original width. We compare the proposed method with four exiting schemes including the uniform scaling, the seam-carving-based video resizing [3], the warping-based video retargeting [4], and the resizing scheme with motion-aware temporal coherence [10]. For subjective performance comparison, readers can obtain the complete set of test results from our project website [26].

A. Performance Evaluation

Fig. 5 shows some snapshots of three videos with different types of camera and object motions resized with the proposed method. The corresponding scaling maps of the nine video frames are shown on the right-hand side. The higher the scaling factor of a pixel, the more visual importance of the pixel. The top row of Fig. 5 shows a video shot with both local (object) and global (camera) motions. The proposed method preserves the foreground object well, and the deformation on the boundary between the object and background regions are visually negligible. This is accomplished because the proposed mosaic-based global scaling map can effectively mitigate the effect of camera motion and object motion.

In the second row of Fig. 5, the test video contains large camera motions. Camera motion is a major factor that causes spatio-temporal incoherence in a resized video. In the energy-based retargeting methods [3], [4], the content scaling is mainly guided by an energy map, but the map is easily influenced by camera motions. Significant camera motions usually lead to large fluctuations between the energy maps of neighboring frames. With large fluctuations in the energy maps, it becomes difficult for the energy-based schemes to maintain the spatio-temporal coherence as the energy values of patches/pixels in objects and background will vary largely as well. The proposed global scaling map serves as a global-motion-compensated temporal filter to smooth out the temporal fluctuation of object size. As shown in Fig. 5, the proposed method generates very smooth, both spatially and temporally, local scaling maps for the three frames even under large camera motions, thereby ensuring the spatio-temporal coherence. Fig. 6 shows the panoramic mosaic of the sequence in the second row of Fig. 5 and the corresponding global scaling map of the mosaic. In the example shown in the bottom row of Fig. 5, the man

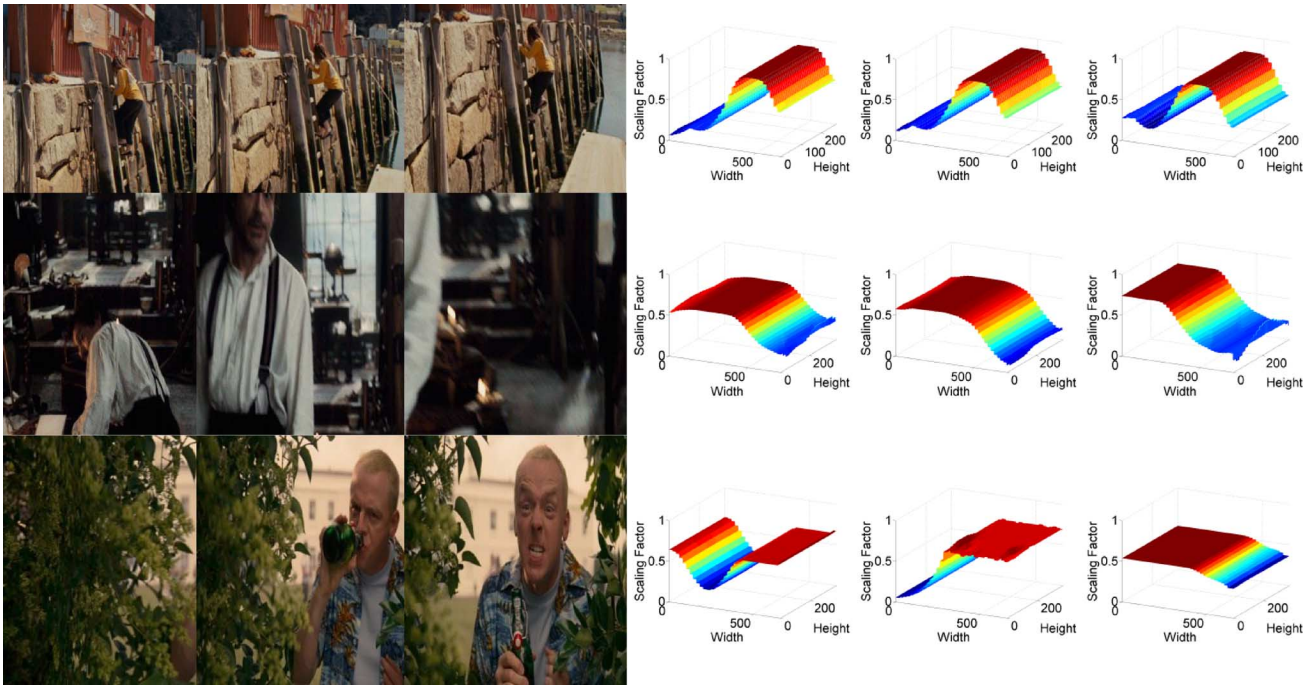

Fig. 5. Snapshots of three videos with different types of camera and object motions resized with the proposed method. The examples show that our method generates smooth scaling maps (shown on the right-hand side) thereby ensuring good spatio-temporal coherence.

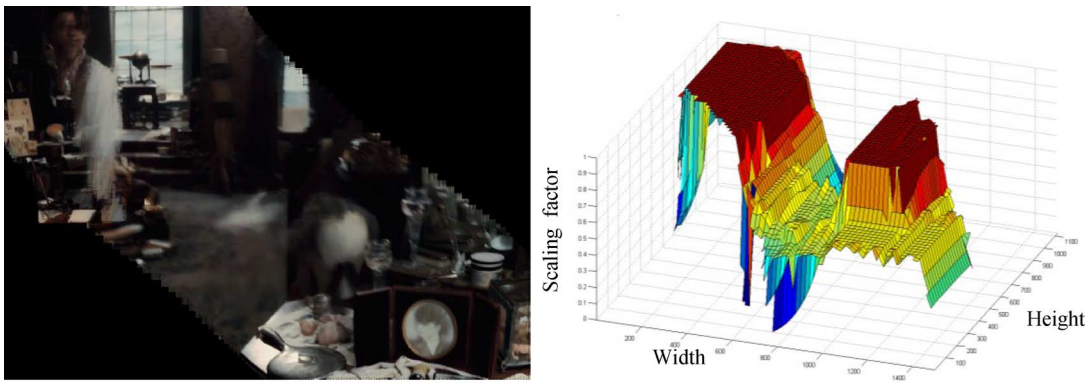

Fig. 6. Panoramic mosaic (shown on the left-hand side) of the sequence in the second row of Fig. 5 and the global scaling map (shown on the right-hand side) of the mosaic.

who appears by the bushes also causes large fluctuations on the corresponding energy maps. Such large fluctuations on the energy maps make traditional saliency-guided resizing schemes fail to effectively maintain the spatio-temporal coherence even by imposing strong spatio-temporal coherence constraints. The proposed method again gracefully maintains the spatio-temporal coherence as it is not sensitive to the fluctuations in the energy maps.

Besides, our proposed mosaic-based method can be easily combined with state-of-the-art image-based retargeting schemes to achieve temporal coherence which the image-based retargeting schemes usually cannot achieve by their own. This is accomplished by using the image-based retargeting method to generate the initial local scaling maps [i.e., to replace the method described in (7)]. The initial scaling maps are then used to generate the global scaling map so as to derive the final local scaling maps. Since the scaling factors of the global scaling map themselves are temporally consistent, the resulting final local scaling maps derived from the global map are also temporally consistent. The temporal coherence can therefore be achieved systematically without the need of introducing content-dependent temporal constraints while solving the scaling allocation problem. For example, we conducted an experiment to integrate the sampling-based method proposed in [6] into our proposed method. To serve this purpose, we first converted the sampling rates of the strips in the sampling-based method into the initial local scaling maps s_L^{ini}. The initial local maps are then fed into the proposed mosaic construction method to generate the global scaling map and the final local scaling maps are obtained accordingly using the process described in Section IV. Fig. 7 compares the performance of the method in [6] with the combination of the method in [6] with the proposed global scaling control method. As illustrated in Fig. 7(a), the sampling-based method obviously leads to inconsistent sizes and shapes of the two foreground objects (the bird and kid) in neighboring frames. The corresponding scaling maps are not

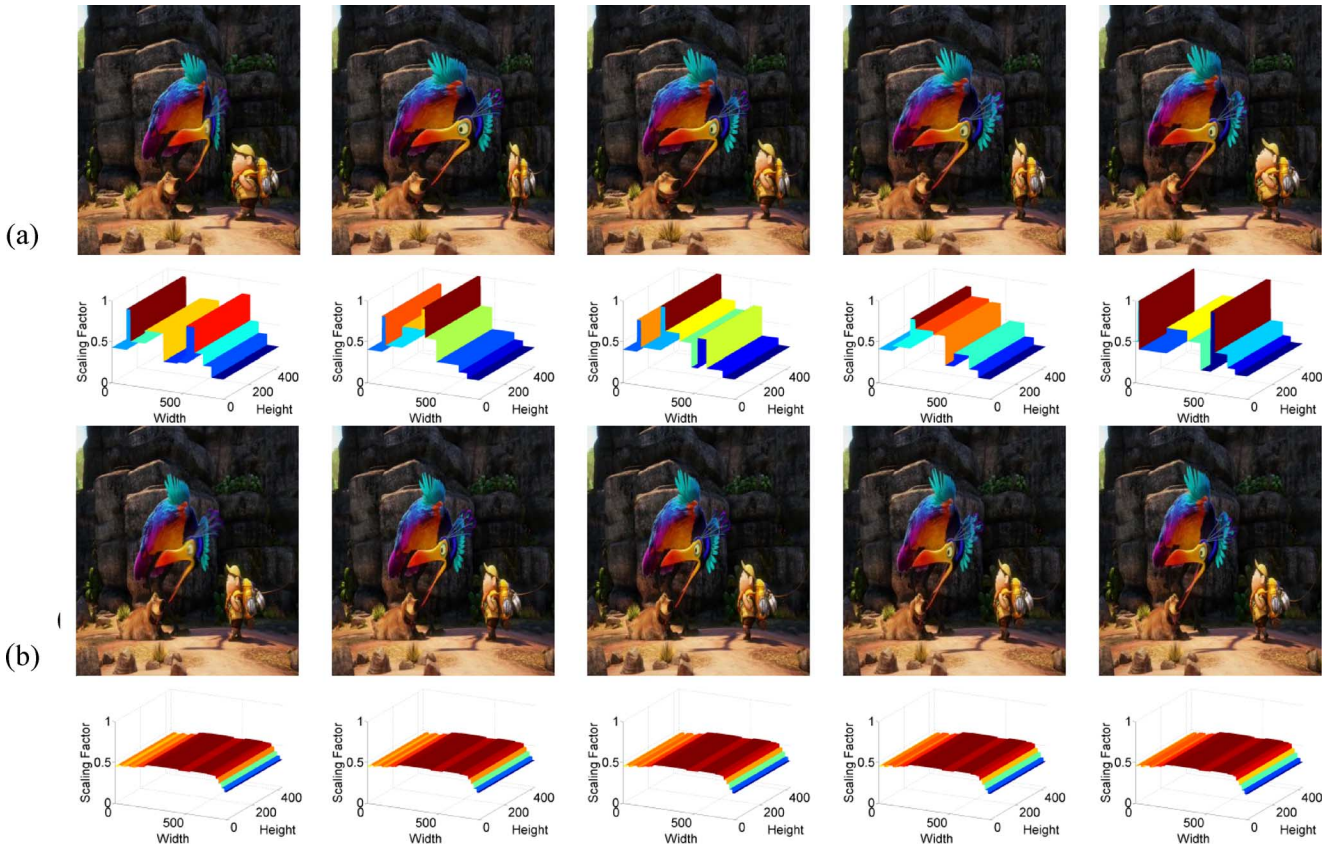

Fig. 7. Performance comparison between the sampling based retargeting method [6] and the combination of the method in [6] with the proposed global scaling control method. From (a), the method in [6] severely distorts the foreground objects. However, after integrating the proposed method, in (b), the retargeting result nicely maintains the temporal coherence and preserves the important content very well.

smooth both temporally and spatially, thereby leading to unpleasant artifacts such as foreground/background deformation and stretching, shrinking, and waving on foreground/background as illustrated in Fig. 7(a) and also in the videos shown in [26]. As can be observed from Fig. 7(b), the proposed method successfully helps the sampling-based method to improve the temporal coherence of the foreground objects.

In Fig. 8, we compare our method with the uniform scaling, seam-carving-based resizing [3], warping-based resizing [4], and Wang *et al.*'s approach [10]. Obviously, uniform scaling is immune to spatio-temporal incoherence distortion caused by any types of camera and object motions. It, however, results in small sized objects and background in important regions. The seam-carving-based video resizing [3] continuously removes a spatio-temporal surface from the video. However, with complex camera and object motions, it is difficult to find low-energy spatio-temporal surfaces that can be removed without significantly distorting the structures of object and background. Therefore the resizing usually causes annoying artifacts. To preserve temporal coherence, the warping-based resizing method [4] constrains the changes of corresponding pixels' positions between neighboring frames. However, it may cause significant visual artifacts when a video contains camera motion or large object motion [see Fig. 8(2d)]. In addition, it also introduces discontinuity when there is an unimportant region in between two important regions. For example, as shown in Fig. 8(3d), the man's hand and the woman's hand are classified as salient regions. The warping-based resizing method, however, introduces severe deformation distortions along the woman's hand.

Wang *et al.* [10] proposed to impose a set of temporal constraints to retain the original object and camera motions as well as to prevent content deformation. Their method blends the aligned saliency map within a sliding window to localize the moving area of an object in the blended saliency map so that the object's size in the moving area can be kept consistent. In this method, the window size cannot be large; otherwise, the blended saliency map will be mostly occupied by moving objects, thereby making it degenerate to the uniform scaling method. However, due to the limited window size for the blended saliency map, the temporal information of video content collected by the method may be too few to generate temporally coherent scaling allocation. As a result, the method may render false camera motion (i.e., shows camera-motion-like effect but there is no camera motion in the original video). Figs. 8(5e) and (6e) show a sequence for that the method proposed in [10] generates the false camera motion artifact (refer to [26]). Furthermore, the method in [10] does not consider the coherence of scaling factors of neighboring patches, which leads to the structure deformation artifact. As shown in Figs. 8(1e) and (2e) where the backgrounds contain several quads, the inconsistent allocation of scaling factors to the quads introduces obvious structure deformations.

Our method was implemented on a personal computer with Duo Core Intel 2.33-GHz CPU and 3-GB memory. Scaling a 320×160 video to 160×160 resolution takes around 0.15

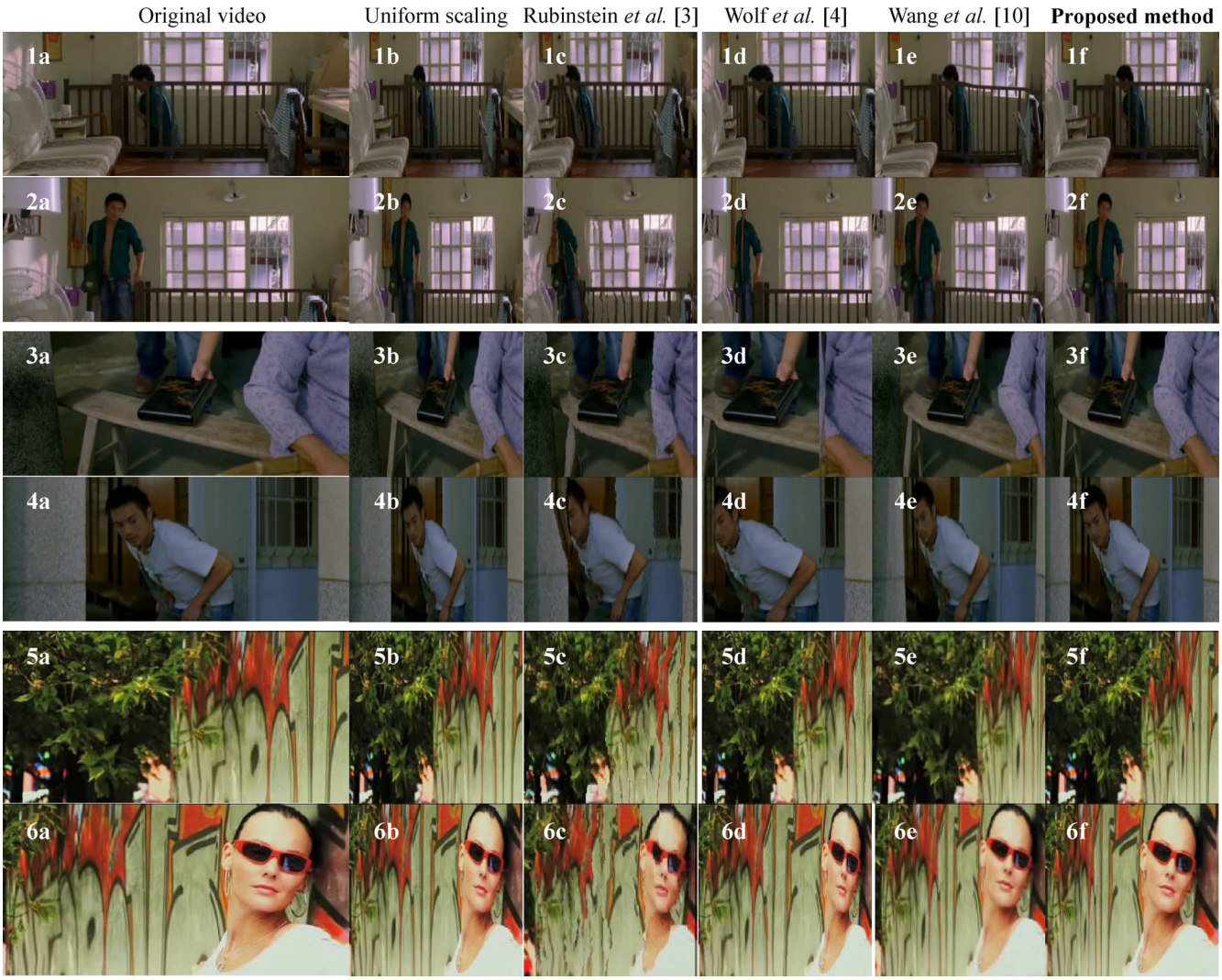

Fig. 8. Subjective comparison of the proposed method with the uniform scaling, the seam-carving-based scaling [3], the warping-based scaling [4], and the video resizing with motion-aware temporal coherence [10]. Our proposed method gracefully preserves the temporal coherence and retains the region-of-interest information.

s/frame to obtain an initial local scaling map (does not include the time of mosaicking) and 2 s/frame to obtain the final local scaling map. Because the shot-level mosaicking consumes the most memory, the memory requirement is dependent on the length of a video shot used for constructing a panoramic mosaic.

B. Limitations

Our method also has its limitations. Like most existing methods, our proposed method will degenerate into the uniform scaling when the frame is mostly occupied with visually important regions (e.g., large-sized objects and complex background), that is, there is very few redundant content that can be removed. Besides, the accuracy of frame alignment will influence the accuracy of global and local scaling maps. In our method, the frame alignment is based on 2-D camera motion estimation which does not consider the distances of feature points to the camera. The simple method may cause misalignment of frames for feature points of different depths. Patch-based schemes are less sensitive to inaccurate frame alignment since a patch's scaling factor is the average of the scaling factors of all pixels in the patch. Because the proposed method can be implemented either in a pixel-based or in a patch-based manner, the requirement on the accuracy of frame alignment can be relaxed when the proposed method is implemented in a patch-based manner.

Currently, our method is more suitable for applications that allow offline processing at the encoder side (e.g., retargeting a prestored video) since its computational complexity is still too high to be used in online video retargeting applications that require real-time processing. Besides, existing mosaicking and foreground/background segmentation tools are still not very mature and reliable for online video processing. However, with offline processing at the encoder side, the constraints would be significantly relaxed. For example, a few interactive video object extraction tools [27], [28] have proven to achieve fairly good and reliable performance for many realistic videos based on user provided rough scribbles labeling the regions of interests. These tools can usually do a good job in offline object segmentation. With successful object segmentation, the difficulty of frame alignment and mosaicking in typical videos would be significantly reduced as well, as long as the background contains enough distinct feature points.

VI. Conclusion

To tackle the spatio-temporal incoherence problem which often occurs in video retargeting, we proposed a novel content-aware video retargeting method for structure-level video adaptation. The proposed method, that is suitable for applications that allow offline processing at the encoder side, is comprised of five major operations: saliency map generation, shot-based panoramic mosaic construction, global scaling map generation, local scaling map generation, and frame resizing. We have presented a constrained energy-preserving optimization method to generate initial frame-level scaling maps based on pixel-wise saliency maps. In addition, we have proposed a mosaic-based global scaling mapping scheme which can systematically maintain temporal coherence of a resized video. The spatial coherence in a frame is further ensured by imposing spatial coherence constraints on the local scaling map of the frame using an iterative optimization manner. Our experimental results show that the proposed method achieves good energy preservation and high spatio-temporal coherence while resizing a video, thereby ensuring good subjective visual quality of the resized video, even when the video contains significant camera motions and object motions. Thanks to its systematic mosaic-based global mapping mechanism, the proposed method can also be easily integrated with other energy-based frame resizing schemes to benefit from the power of existing or new frame resizing tools.

Acknowledgment

The authors would like to thank Prof. B. Sankur and the anonymous reviewers for their valuable comments that help improve the quality of this paper. They would also like thank M. Rubinstein and Dr. Y.-S. Wang for their suggestions about experiment setup and for providing some simulation results based on the methods in [3] and [10], and Prof. W. Lin for providing the software of PQSM [20].

References

[1] A. Shamir and O. Sorkine, "Visual media retargeting," in *Proc. ACM SIGGRAPH ASIA Courses (SIGGRAPH ASIA'09)*, 2009, pp. 1–13.

[2] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 10:1–10:10, Aug 2007.

[3] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 16:1–16:9, Aug. 2008.

[4] 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.

[5] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 126:1–126:10, Dec. 2009.

[6] J.-S. Kim, J.-H. Kim, and C.-S. Kim, "Adaptive image and video retargeting technique based on fourier analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Kyoto, Japan, Sep. 2009, pp. 1730–1737.

[7] L. Shi, J. Wang, L. Y. Duan, and H. Lu, "Consumer video retargeting: Context assisted spatial-temporal grid optimization," in *Proc. ACM Int. Conf. Multimedia*, Beijing, China, Oct. 2009, pp. 301–310.

[8] S. Kopf, T. Haenselmann, D. Farin, and W. Effelsberg, "Automatic generation of summaries for the web," in *Proc. IS&T/SPIE Conf. Storage Retrieval For Media Databases*, San Jose, CA, Jan. 2004, pp. 417–428.

[9] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, Oct. 2006, pp. 241–250.

[10] Y.-S. Wang, H. Fu, O. Sorkine, T.-Y. Lee, and H.-P. Seidel, "Motion-aware temporal coherence for video resizing," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 127:1–127:10, Dec. 2009.

[11] W.-H. Cheng, C.-W. Wang, and J.-L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.

[12] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, "Retargeting images and video for preserving information saliency," *IEEE Comput. Grap. Appl.*, vol. 27, no. 5, pp. 80–88, Sep./Oct. 2007.

[13] Y. Guo, F. Liu, J. Shi, Z.-H. Zhou, and M. Gleicher, "Image retargeting using mesh parametrization," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 856–867, Aug. 2009.

[14] T. Ren, Y. Liu, and G. Wu, "Image retargeting based on global energy optimization," in *Proc. IEEE Int. Conf. Multimedia Expo*, New York, Jun. 2009, pp. 406–409.

[15] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 118:1–118:8, Dec. 2008.

[16] C.-K. Chiang, S.-F. Wang, Y.-L. Chen, and S.-H. Lai, "Fast JND-based video carving with GPU acceleration for real-time video retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 11, pp. 1588–1597, Nov. 2009.

[17] 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.

[18] 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.

[19] Y.-F. Zhang, S.-M. Hu, and R. R. Martin, "Shrinkability maps for content-aware video resizing," *Comput. Graph. Forum*, vol. 27, no. 7, pp. 1797–1804, 2008.

[20] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1928–1942, Nov. 2005.

[21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[22] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[23] R. Szeliski, "Image alignment and stitching: A tutorial," *FNT Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, 2006.

[24] 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.

[25] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J Optim*, vol. 9, no. 4, pp. 877–900, 1999.

[26] NTHU, Video Scaling Project. [Online]. Available: http://www.ee.nthu.edu.tw/cwlin/scaling/scaling.htm

[27] Y. Li, J. Sun, and H.-Y. Shum, "Video object cut and paste," *ACM Trans. Graph.*, vol. 24, pp. 595–600, 2005.

[28] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: Robust video object cutout using localized classifiers," *ACM Trans. Graph.*, vol. 28, pp. 70:1–70:12, Jul. 2009.

Tzu-Chieh Yen received the B.S. degree from National Central University, Taoyuan, Taiwan, in 2008, and the M.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 2010, both in electrical engineering.

He has been with ASUSTeK Computer Inc., Taipei, Taiwan, as a Design Engineer since September 2010. His research interests include video coding and video content adaptation.

Chia-Ming Tsai received the B.S. degree from Feng Chia University, Taichung, Taiwan, in 2003, and the M.S. degree from National Chung-Cheng University, Chiayi, Taiwan, in 2005, both in computer science and information engineering. He is currently working toward the Ph.D. degree in the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan.

His research interests include video coding and video content adaptation.

Chia-Wen Lin (S'94–M'00–SM'04) received the 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, NTHU. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University (CCU), Taiwan, from 2000 to 2007. Prior to joining academia, he worked for the Information and Communications Research Laboratories, Industrial Technology Research Institute (ICL/ITRI), Hsinchu, Taiwan, from 1992 to 2000, where his final post was Section Manager. From April 2000 to August 2000, he was a Visiting Scholar with Information Processing Laboratory, Department of Electrical Engineering, University of Washington, Seattle. He has authored or coauthored over 90 technical papers. He holds more than 20 patents. His research interests include video content analysis and video networking.

Dr. Lin is Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Journal of Visual Communication and Image Representation*. He has served as a Guest Co-Editor of four special issues for the IEEE TRANSACTIONS ON MULTIMEDIA, the *EURASIP Journal on Advances in Signal Processing*, and the *Journal of Visual Communication and Image Representation*, respectively. He served as Technical Program Co-Chair of the IEEE International Conference on Multimedia and 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 SPIE 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.