Mosaic-guided Video Retargeting for Video Adaptation

Chia-Ming Tsai[a], Tzu-Chieh Yen[b], and Chia-Wen Lin*[c]
[a] Dept. of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 62102, Taiwan;
[b] ASUSTek Computer Inc., Taipei 11259, Taiwan
[c] Dept. of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan

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. In this paper, we propose a mosaic-guided video retargeting scheme to ensure good spatio-temporal coherence of the downscaled video. Besides, a rate-distortion optimization framework is proposed to maximize the information retained in the downscaled video.

Keywords: Video Adaptation, video retargeting, video scaling, spatio-temporal coherence

1. 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. No matter how the visual content is resized to another lower resolution, it cannot prevent information loss from its full-resolution version.

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 2.

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 consistency of neighboring frames, leading to variation of the scaling factor of a corresponding region in neighboring frames. Such inconsistency leads to 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.

The proposed method is an extended version of our previous work [29]. Our primary goal is to solve the temporal incoherence problem in a systematic way, rather than resorting to numerous temporal coherence constraints. To ensure good temporal coherence, our proposed method first constructs a panoramic mosaic for a video shot. Besides, in order to keep scaling factor coherence inside each object, we adopt a semi-automatic object segmentation method to identify the object regions in the panorama mosaic. Based on the object masks, the mapping relationships of individual frames to the

panorama mosaic map are used to generate scaling factor constraints. By imposing the constraints on scaling factors and on scaling factor variation ins each object regions, the proposed method directly resizes the panoramic mosaic to obtain the optimized scaling maps of individual frames. Consequently, the scaling maps of individual frames are derived according to their mapping relationships to the panoramic mosaic. The proposed method avoids the iterative optimization procedure for individual frames by translating the video retargeting problem into an image retargeting problem under scaling factor constraints.

The rest of this paper is organized as follows. Section 2 summarizes the state-of-the-art content-aware video retargeting approaches. Our proposed mosaic-guided scaling method is presented in Section 3. Section 4 reports and discusses the experimental results. Finally, conclusions are drawn in Section 5.

2. 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. Accordingly, 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 re-compute 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 3D rectilinear grid resizing scheme. The performance of the method was evaluated on sports and advertisement videos. The cropping-based methods proposed in [8], [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.

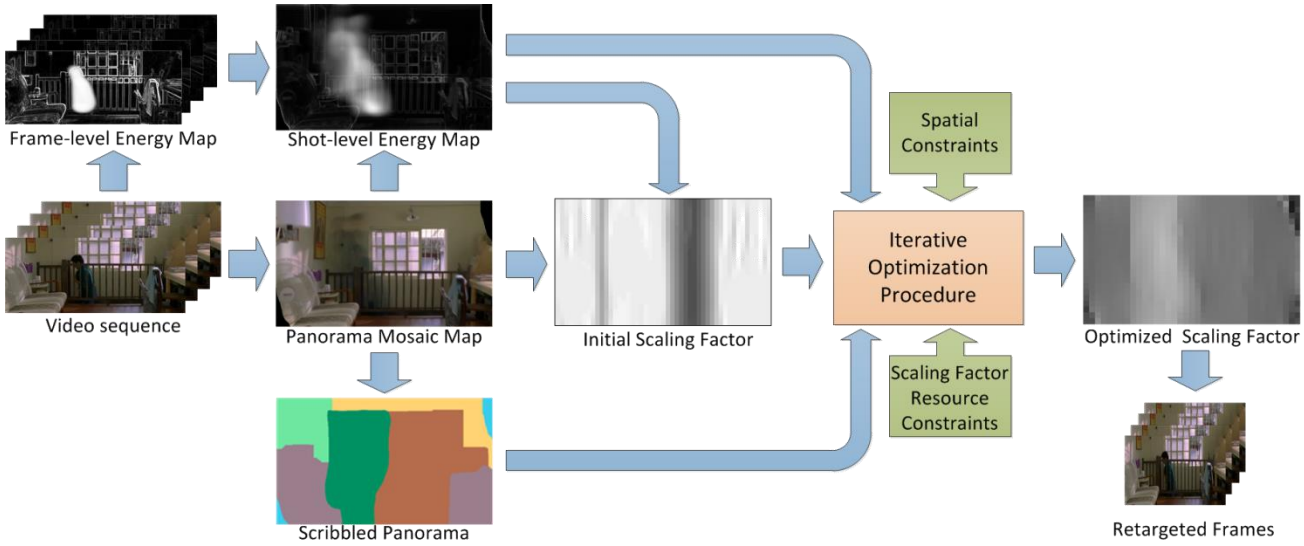

Figure 1. Flow diagram of the proposed method.

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. Our previous work [29] proposed to construct a panorama mosaic for a video shot to keep spatio-temporal coherence in the shot. The panoramic mosaic is 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. However, the method proposed in [29] requires an iterative optimization procedure to derive the local scaling maps of individual frames, which is time consuming and the resizing result is sensitive to the influence of sudden saliency change.

Different from the existing schemes, our proposed mosiac-guided scaling scheme is a hybrid approach. Our scheme constructs a panoramic mosaic from a spatio-temporal cube (e.g., a video shot) to record the object and camera motions. The proposed method then directly resize the panorama mosaic map to derive the optimal scaling maps of individual frames. This is achieved by imposing the avalable scaling budget constraints in the panorama mosaic map. As a result, the new retargeting approach makes a more robust global decision of scaling factors so as to mitigate the influence of object and camera motions.

3. PROPOSED VIDEO RETARGETING SCHEME

Assume we resize a video from resolution $W \times H$ 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 $\mathbf{I}_{\text{in}} = \left\{ I_{\text{in}}^{(t)} \right\}_{t=1}^{N}$ and the corresponding resized frames are denoted as $\mathbf{I}_{\text{out}} = \left\{ I_{\text{out}}^{(t)} \right\}_{t=1}^{N}$. Video retargeting is to find a transform $I_{\text{out}}^{(t)} = \mathbf{T}\left(I_{\text{in}}^{(t)} \right)$ which can preserve in the resized frame the most important content while maintaining spatio-temporal coherence. As illustrated in Fig. 1, the proposed method involves six major operations to tackle the video retargeting problem: energy map generation, shot-level panoramic mosaic construction, semi-automatic object segmentation, initial scaling map generation, iterative constrained optimizatione, and frame resizing. The detailed operations of the proposed retargeting scheme are elaborated below.

3.1.1 Initialization

The proposed mosaic-guided scaling method needs four kinds of maps for resizing a video shot: the frame-level energy maps, the shot-level panoramic mosaic, the shot-level energy map, and the object masks.

3.1.2 The 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. 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

$$e(i, j) = \alpha_1 \cdot Gradient(i, j) + \alpha_2 \cdot PQSM(i.j) \tag{1}$$

where $e(i,j)$ represents the energy value of the (i,j)-th pixel. The values of $Gradient(i,j)$ and $PQSM(i,j)$ 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].

3.1.3 The 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 [21] 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 [21]. 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 [22] 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 (empirically set as 0.6), 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.

3.1.4 The Shot-Level Energy Map

In order to obtain the initial scaling factor value, we adopt a linear programming solver to solve the constrained optimization problem to obtain the initial scaling maps. The solver takes a shot-level energy map as the guide to derive the initial scaling maps. Therefore, the shot-level energy map is generated by fusing every frame-level energy maps based on the corresponding locations inside the panorama mosaic map.

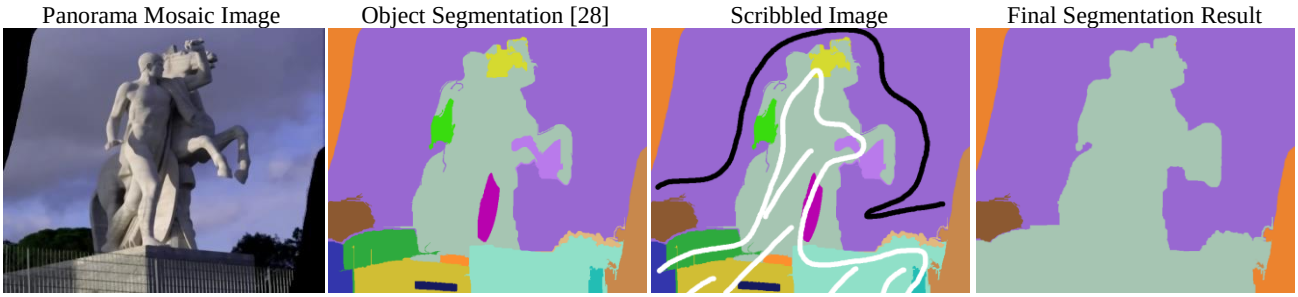

Figure 2. An example of object segmentation for the panoramic mosaic image via a semi-automatic object segmentation method. (a) Panorama mosaic image; (b) Automatic object segmentation result [28]; (c) User scribbled image; (d) The final segmentation result.

Let $\mathbf{H}^{(t)}$ denote the projective transform of the *t*-th frame, $(i, j)_{\text{in}}^{(t)}$ the coordinate of the (*i*,*j*)-th pixel in the *t*-th 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 shot-level energy map is simply obtained as the union of energy values after the transformation, as expressed by

$$e\left((i', j')_M\right) = \bigcup_t \bigcup_{(i,j)_{in}^{(t)} \to (i', j')_M} e\left(\mathbf{H}^{(t)}(i, j)_{\text{in}}^{(t)}\right), \tag{2}$$

where $e\left((i', j')_M\right)$ represents a set of energy values corresponding to pixel (i', j') of the panoramic mosaic. Note, the union operation in (2) is a many-to-one mapping, that is, $(i', j')_M$ may correspond to the energy values from different video frames and different pixels of the original video.

To obtain a single-valued mapping, we choose the mean energy value in the set defined in (2) as the energy value for pixel $(i', j')_M$ of the global scaling map as follows:

$$e_G\left((i', j')_M\right) = \text{mean}\left\{e\left((i', j')_M\right)\right\}. \tag{3}$$

3.1.5 The Semi-automatic object segmentation

Since the scaling factors of pixels/patches in an object should be kept consistent, we adopt an object segmentation tool to identify objects. However existing automatic object segmentation tools are still not very mature and reliable and may have over-segmentation or under-segmentation problems. To avoid these problems, we let user can participate in the segmentation process. As illustrated in Fig. 2(b), the segmentation result by the automatic segmentation tool proposed in [28] still has over-segmentation problem. Therefore, the user can scribble the segmentation image with different colors to merge the regions that belong to the same object. Fig. 2(c) shows the user scribbled image and Fig. 2(d) shows the final object segmentation result.

Although the proposed method uses a semi-automatic object segmentation tool to segment all objects in the panorama mosaic image, it still can be used in on-line retargeting system. The segmentation can be performed off-line, and the segmentation masks cane be stored as metadata. The metadata are subsequently used to significantly reduce computation while performing on-line retargeting at the encoder/decoder, thereby relaxing the complexity constraint as well as achieving a good tradeoff between visual quality, format flexibility, and on-line complexity.

3.2 Mosaic-Guided Video Retargeting

The scaling factor change between the resized frames should be constrained by fitting the mapping model. To this end, a panoramic mosaic is constructed for a video shot to maintain the temporal coherence of video resizing under camera and object motions. The initial scaling maps of frames in the video shot then derived from the panoramic mosaic. We then perform an iterative optimization procedure to generate the optimal scaling maps of individual frames based on the initial scaling maps, scaling budget constraints, and spatial coherence constraints. In this section, we first introduce the method of generating the initial scaling maps and then present the iterative constrained optimization process of generating the scaling maps of individual frames.

3.2.1 The Initial Scaling Factor

Our method uses an energy-based frame resizing approach to determine the initial scaling factor values. The initial scaling map of panorama mosaic is obtained by solving a constrained energy-preserving optimization problem that is to maximize the energy retained in a resized panorama mosaic by

$$s_G^{ini*} = \arg\max_{\left\{s_G^{ini}(i',j')\right\}} \sum_{j'=1}^{H_G}\sum_{i'=1}^{W_G} e_G\left((i',j')_M\right)\cdot s_G^{ini}\left((i',j')_M\right)$$
$$\text{s.t. } \sum_{i'} s_G^{ini}\left((i',j')_M\right) = W', \text{ and } \left|s_G^{ini}\left((i',j')_M\right) - s_G^{ini}\left((i',j'+1)_M\right)\right| \le TH_s, \forall j \tag{4}$$

where s_G^{ini*} represents the optimal initial scaling factor map, $e_G\left((i',j')_M\right)$ denotes the energy magnitude at pixel $(i',j')_M$ of the saliency map. $s_G^{ini}\left((i',j')_M\right)$ denotes the initial local scaling factor for pixel $(i',j')_M$ in the panorama mosaic, where $0 \le s_G^{ini}\left((i',j')_M\right) \le 1$, and W' is the target width. The threshold for the imposed spatial constraint $TH_s = 0.06$ for all input videos.

3.2.2 Scaling Budget Constraints

The total available scaling budget is limited when resizing an image. For example, suppose there is an image line of size 800x1. If the line is downscaled by two, no matter how we adjust scaling factors of pixels on the line, the resized line length is constrained to 400. Therefore, we can resize a video by directly resizing the panorama image according to the available scaling budget. As mentioned in Section 3.1.3, the projection of a coordinate is given by $(i',j')_M = \mathbf{H}^{(t)}(i,j)_{\text{in}}^{(t)}$. Because we use a simplified affine model with only scaling and translation parameters, the horizontal scaling factor resource constraint form the *t*-th frame is given as follows:

$$\sum_{i'} S_G\left((i',j')_M\right) = \left|\left(p_1^{(t)},j'\right)_M - \left(p_2^{(t)},j'\right)_M\right| \times \frac{W'}{W}, \forall j', \tag{5}$$

where $\left(p_1^{(t)},j'\right)_M = \mathbf{H}^{(t)}(1,j)_{\text{in}}^{(t)}$ and $\left(p_2^{(t)},j'\right)_M = \mathbf{H}^{(t)}(W,j)_{\text{in}}^{(t)}$ denote the left and right coordinates of the corresponding positions of the *t*-th frame. Similarly, the vertical scaling factor resource constraint form the *t*-th frame is given as follows:

$$\sum_{j'} S_G\left((i',j')_M\right) = \left|\left(i',q_1^{(t)}\right)_M - \left(i',q_2^{(t)}\right)_M\right| \times \frac{H'}{H}, \forall i', \tag{6}$$

where $\left(i',q_1^{(t)}\right)_M = \mathbf{H}^{(t)}(i',1)_{\text{in}}^{(t)}$ and $\left(i',q_2^{(t)}\right)_M = \mathbf{H}^{(t)}(i,H)_{\text{in}}^{(t)}$ are the upper and lower coordinates of the corresponding positions of the *t*-th frame.

3.2.3 Information Loss Constraint

A video retargeting method should avoid critical visual information loss. The information loss after resizing the panorama mosaic map can be measured by the energy distortion between the original image and the resized one as follows:

$$D_{Info} = \sum_{j'=1}^{H_G} \sum_{i'=1}^{W_G} \left(1 - e_G\left(\left(i', j'\right)_M\right)\right) \cdot S_G^{(n)}\left(\left(i', j'\right)_M\right) \tag{7}$$

$e_G\left(\left(i', j'\right)_M\right)$ and $S_G^{(n)}\left(\left(i', j'\right)_M\right)$ respectively represent the energy value and the *n*-th round scaling factor of pixel $\left(i', j'\right)_M$ in the panorama mosaic map, and $0 \le S_G^{(n)}\left(\left(i', j'\right)_M\right) \le 1$.

3.2.4 Spatial Coherence Constraints

In the optimization process, we impose the following constraints to prevent the spatial incoherence distortion.

1) **Object Deformation Constraint.** 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 corresponding patches/pixels in neighboring frames may have different spatial locations, sizes, and shapes, thereby being scaled differently. Such inconsistent scaling for corresponding patches/pixels in neighboring frames results in temporal incoherence artifacts such as stretching, shrinking, and waving of object or background. To prevent the temporal artifacts, the scaling factors of the same visual content should be kept as consistent as possible in neighboring frames. Besides, to maintain spatial coherence, the scaling factors within each object should also be made consistent. To do so, we define a set $\mathbf{O} = \{O_1, O_2, O_3, ..., O_K\}$ consisting of all objects in the panorama mosaic map, where *K* is the number of extracted objects. To maintain the consistency of each object size, the following spatial scaling inconsistency distortion should be minimized:

$$D_{SO} = \sum_n \sum_{(i', j')_M \in O_k} \left| S_G^{(n)}\left(\left(i', j'\right)_M\right) - S_G^{(n)}\left(O_k\right) \right| \tag{8}$$

$$S_G^{(n)}\left(O_k\right) = \frac{1}{A_{O_n}} \times \sum_{(i', j')_M \in O_k} S_G^{(n)}\left(\left(i', j'\right)_M\right) \tag{9}$$

where $S_G^{(n)}\left(O_k\right)$ denotes the *n*-th round average scaling factor of object O_n. In brief, we minimize the variation of scaling factors in each object to maintain the consistency of each object size.

2) **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 as follows:

$$D_{SS} = \sum_{j'} \sum_{i'} \left| s_G^{(n)}\left(\left(i', j'\right)_M\right) - s_G^{(n)}\left(\left(i', j'+1\right)_M\right) \right|. \tag{10}$$

3.2.5 Iterative Optimization Procedure

After obtaining the first round scaling factor map $s_L^{(1)} = s_G^{ini*}$, an iterative optimization procedure is performed to find a converged solution s_G^* subject to three smoothness constraints: (7), (8), and (10). The final refined scaling maps of individual frames are derived iteratively from (11) using an iterative optimization solver.

$$s_G^* = \arg\min_{s_G^{(n)}} D_{\text{total}} = \arg\min_{s_G^{(n)}} \left(D_{Info} + \lambda_1 D_{SO} + \lambda_2 D_{SS} \right). \tag{11}$$

$$\text{s.t. } \left| s_G^{(n)}\left((i', j')_M\right) - s_G^{(n)}\left((i', j'+1)_M\right)\right| \le TH_{SS}, \quad \forall i', j'$$

$$\sum_{i'} S_G\left((i', j')_M\right) = \left|\left(p_1^{(t)}, j'\right)_M - \left(p_2^{(t)}, j'\right)_M\right| \times \frac{W'}{W}, \forall j'$$

where λ_1 and λ_2 are the weighting factor for D_{SO} and D_{SS}, respectively. In our implementation, we set D_{Info}, D_{SO}, and D_{SS} equally important (i.e., $\lambda_1 = \lambda_2 = 1$). The threshold for spatial constraint TH_{SS}, similar to the case in (4), is empirically set to be 0.06.

3.2.6 Frame Resizing Based on the Optimized Scaling Factor Map

After obtaining the optimal local scaling map of a frame, the frame is scaled accordingly. When the video has room-in/out effect, the mapped size of each frame may not be the same as its original frame size. As a result, after getting the scaling factor from s_G^*, it needs to apply the inverse mapping, $\mathbf{H}'^{(t)}$, to derive the final local scaling map so as to fit the target video size as given below:

$$S_L^*(i, j) = S_G^*\left(\mathbf{H}'^{(t)}\left(i', j'\right)_M\right) \tag{12}$$

After obtaining the final local scaling maps, the resized frame is generated by the pixel fusion based image downscaling proposed in [13]. 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.

4. 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 and drama videos. In the experimental settings, each test video is resized to the half size of the original width. We compare the proposed method with three exiting schemes including the uniform scaling the resizing scheme with motion-aware temporal coherence [9], our previous work [29]. For subjective performance comparison, readers can obtain the complete set of test results from our project website [24].

4.1 Performance Evaluation

First, to evaluate the impact of object segmentation on maintaining the consistency of object size, we compare the global scaling maps obtained with the proposed method and with our previous method [29]. Fig. 3(a) to Fig. 3(c) illustrate the panorama mosaic image, the shot-level energy map; and the semi-automatic object segmentation result, respectively. As can be observed in Fig. 3(d) and Fig. 3(e), the proposed method successfully keeps the coherence of each object size compared to our previous method [29] that may over-trim some areas in the background region. The proposed method uses semi-automatic method to separate objects of the panorama mosaic image. It not only mitigates the artifacts due to camera and object motions but also preserves the consistency of object size after resizing. Besides, the method in [29] is sensitive to sudden saliency change, which might lead to slight false-shift artifacts. In contrast, the proposed method derives the optimal scaling factor by directly retargeting the panorama mosaic image, making the global decision procedure more robust and thereby eliminating the false-shift artifact.

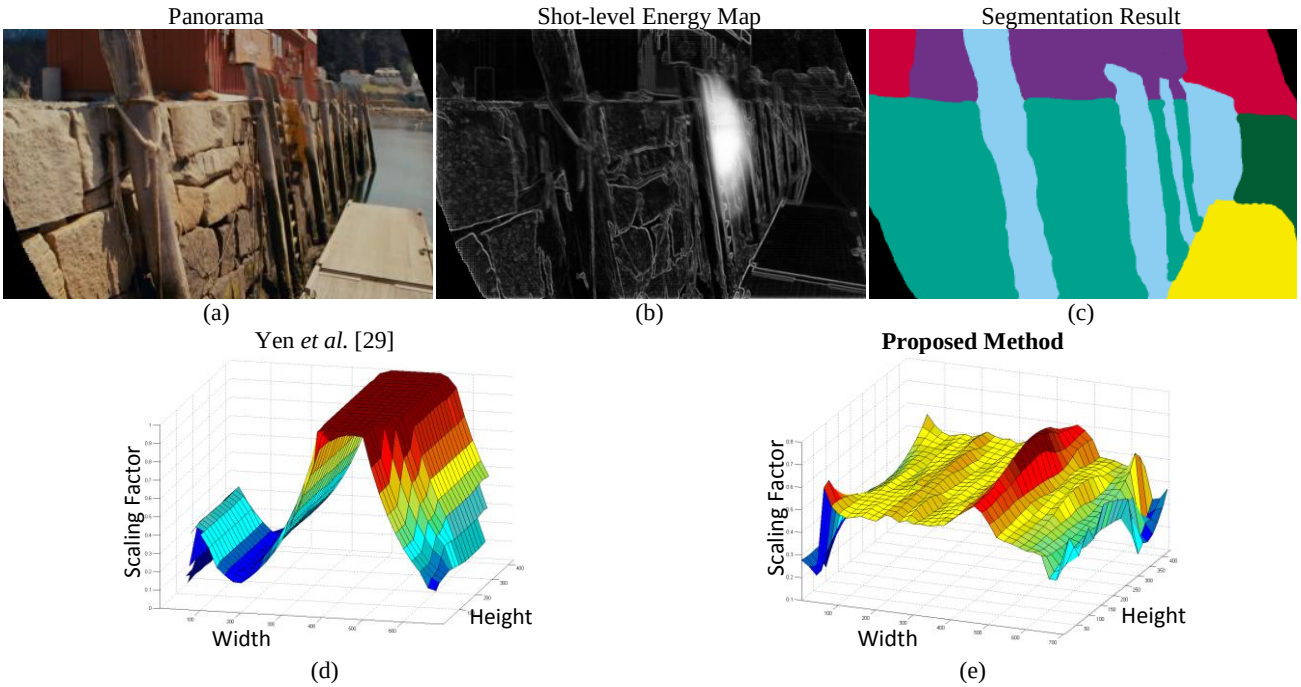

Figure 3. Comparison of scaling factor maps: (a) Panorama mosaic image; (b) Shot-level energy map; (c) Segmentation result; (d) The global scaling map obtained with the method in [29]; (e) The global scaling map of obtained with the proposed method.

In Fig. 4, we compare our method with the uniform scaling, Wang *et al.*'s approach [10], and our previous work [29]. 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. 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). Fig. 4(7c) and Fig. 4(8c) show a sequence for that the method proposed in [10] generates the false camera motion artifact (refer to [25]). 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 Fig. 4(1c) and Fig. 4(2c) where the backgrounds contain several quads, the inconsistent allocation of scaling factors to the quads introduces obvious structure deformations.

Our previous work [29] separates each frame into foreground and background. However if there are visually important areas in background, it may cause uncomfortable artifacts. Fig. 4(1d) and Fig. 4(2d) show the artifact on the right-hand-side desk. Besides, the resizing result might cause slight false-shifting artifact. The main cause is, to satisfy the boundary constraints in a frame, the scaling factors for the same object between adjacent frames may be different, thereby causing unnatural artifact in the temporal domain. In contrast, the proposed method can meet both the scaling factor resource constraint and the boundary constraints in each frame by evaluating the scaling factor only once, so as to avoid the false-shifting artifact.

Our method was implemented on a personal computer with Intel Core 2 Quad Q6600 CPU and 6 GB memory. For a 320x160 test sequence with 184 frames, scaling the video to 160x160 resolution takes around 3 seconds to obtain the initial scaling factor map (does not include the time of mosaicking) and 251 seconds (1.36 sec/frame) to derive the optimized scaling factor map. Because the shot-level mosaicking consumes most memory, memory requirement is dependent on the length of a video shot used for constructing a panoramic mosaic.

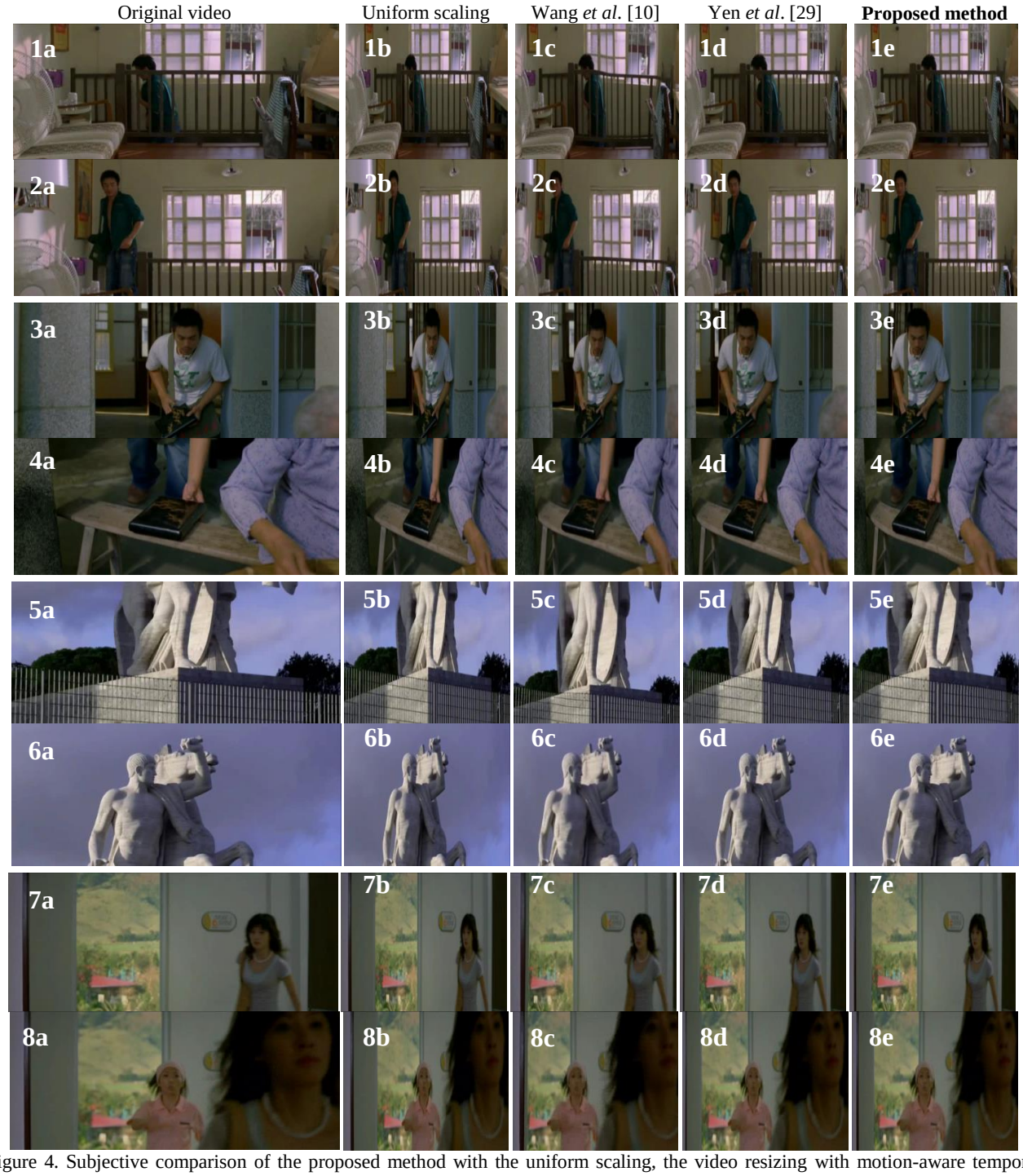

Figure 4. Subjective comparison of the proposed method with the uniform scaling, the video resizing with motion-aware temporal coherence [10], our previous work [29], and the proposed method.

4.2 Limitations

Our method also has its limitations. The proposed method uses the semi-automatic object segmentation method to segment the panorama mosaic image into many object regions. When the visual importance of different regions varies significantly, the proposed method tends to keep the size of importance regions while over-trimming the other

unimportant regions, leading to uncomfortable visual artifact. Fig. 5 shows an example of the situation. As can be observed in Fig. 5(b), the energy value of the arch rock is intentionally emphasized in the panorama mosaic image. Our method preserves the content of the arch rock and trims the regions at left and right sides. The resized result looks not nature and does not keep the room-out effect of the video well. Besides, the accuracy of frame alignment will influence the accuracy of final scaling factor values and scaling factor resource constraints. In our method, the frame alignment is based on 2D 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.

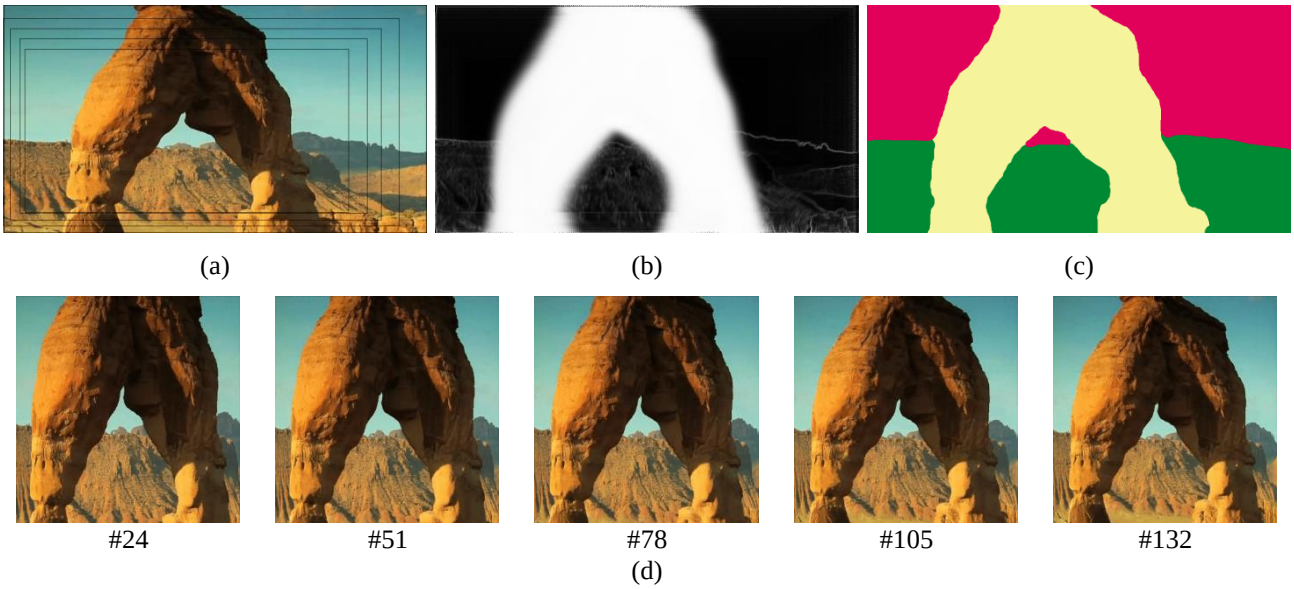

Figure 5. A video retargeting example when the visual importance of each region is too different. (a) Panorama mosaic image; (b) Energy map that highlights the energy of the arch rock; (c) User defined object regions; (d) Snapshots of resized frames (frames #24, #51, #78, #105, and #132).

5. 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, is comprised of six major operations: energy map generation, shot-level panoramic mosaic construction, semi-automatic object segmentation, initial scaling map generation, scaling map refinement, and frame resizing. We have presented a constrained energy-preserving optimization method to generate initial scaling maps based on panorama mosaic and shot-level saliency map. Besides, we have proposed a mosaic-based global scaling mapping scheme which can systematically maintain temporal coherence of a resized video. The spatial coherence in each frame is further ensured by imposing scaling factor resource constraints on the scaling map refinement procedure. 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.

6. REFERENCES

[1] A. Shamir and O. Sorkine, “Visual media retargeting,” in *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. Graphics*, vol. 26, no. 3, pp. 16, 2007.

[3] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Trans. Graphics*, vol. 27, no. 3, pp. 16, 2008.

[4] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1−6, Rio de Janeiro, Brazil.
[5] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graphics*, vol. 28, no. 5, 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.*, pp. 1730−1737, Sept. 2009, Kyoto, Japan.
[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*, pp. 301−310, Oct. 2009, Beijing, China.
[8] S. Kopf, T. Haenselmann, D. Farin, and W. Effelsberg, " Automatic generation of summaries for the Web," in *Proc. IS&T/SPIE Conf. Storage and Retrieval for Media Databases*, pp. 417–428, Jan. 2004, San Jose, USA.
[9] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," in *Proc. ACM Int. Conf. Multimedia*, pp. 241−250. Oct. 2006, Santa Barbara, CA.
[10] Y.-S. Wang, H. Fu, O. Sorkine, T.-Y. Lee, and H.-P. Seidel, "Motion-aware temporal coherence for video resizing," *ACM Trans. Graphics*, vol. 28, no. 5, 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 Computer Graphics and Applications*, vol. 27, no. 5, pp. 80−88, Sept.–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*, pp. 406−409. June 2009, New York, USA.
[15] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graphics*, vol. 27, no. 5, pp. 118, 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.*, pp. 1026−1033, 2009, Kyoto, Japan.
[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*, pp. 321−330, Oct. 2009, Beijing, China.
[19] Y.-F. Zhang, S.-M. Hu, and R. R. Martin, "Shrinkability maps for content-aware video resizing," *Computer Graphics 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] 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," *Foundations and Trends in Computer Graphics and Vision (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, June 1981.
[24] R .H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optimization*, vol. 9, no. 4, pp. 877–900, 1999.
[25] NTHU Video Scaling project. [Online]. Available: http://www.ee.nthu.edu.tw/cwlin/scaling/scaling.htm.
[26] Y. Li, J. Sun, and H.-Y. Shum, "Video object cut and paste," *ACM Trans. Graphics*, vol. 24, pp. 595−600, 2005.
[27] X Bai, J Wang, D Simons, and G Sapiro, "Video SnapCut: Robust video object cutout using localized classifiers," *ACM Trans. Graphics*, vol. 28, 2009.
[28] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *Int. J. Comput. Vis.*, Vol. 59, Num. 2, Sept. 2004.
[29] 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.