Content-aware Video Retargeting Using Object-preserving Warping

Shih-Syun Lin, Chao-Hung Lin, *Member, IEEE*, I-Cheng Yeh, Shu-Huai Chang, Chih-Kuo Yeh, and Tong-Yee Lee, *Senior Member, IEEE*

Abstract—A novel content-aware warping approach is introduced for video retargeting. The key to this technique is adapting videos to fit displays with various aspect ratios and sizes while preserving both visually salient content and temporal coherence. Most previous studies solve this spatiotemporal problem by consistently resizing content in frames. This strategy significantly improves the retargeting results, but does not fully consider object preservation, sometimes causing apparent distortions on visually salient objects. We propose an object-preserving warping scheme with object-based significance estimation to reduce this unpleasant distortion. In the proposed scheme, visually salient objects in 3D space-time space are forced to undergo as-rigid-as-possible warping, while low-significance contents are warped as close as possible to linear rescaling. These strategies enable our method to consistently preserve both the spatial shapes and temporal motions of visually salient objects, and avoid over-deformations on low-significance objects, yielding a pleasing motion-aware video retargeting. Qualitative and quantitative analyses, including a user study and experiments on complex videos containing diverse cameras and dynamic motions, show a clear superiority of our method over related video retargeting methods.

Index Terms—Video retargeting, spatial and temporal coherence, optimization, warping

1 Introduction

MESH and Seam carving are recent techniques of content-aware retargeting. Seam carving iteratively removes a surface seam passing through insignificant regions, and video warping optimizes the mapping from a source to a target video using various spatial and temporal constraints. While these two techniques generate good results, it should be noted that seam carving may yield jagged edges due to the removal of discontinuous seams, and video warping (i.e., a continuous solution) may generate motion distortions due to the inconsistent scaling factors of grids occupied by an object [1]. To reduce motion distortion in resizing, an object-preserving warping is proposed. The basic idea behind our method is to measure content significance and resize videos by utilizing information of object motions rather than pixel motions as adopted in the previous studies [2], [3], [4]. Information of object motions implies inter-frame object correspondence, which allows the definition of a significance map for volumetric objects and the consistent preservation of both the shapes and motions of volumetric objects in warping.

In the proposed retargeting scheme, the frames in the original video are consistently segmented into several patches in preprocessing, and the corresponding patches in frames are assigned the same significance value in consideration of consistent warping and temporal coherence preservation. Volumetric objects, i.e., a sequence of corresponding patches in frames, with high significance values are forced to undergo as-rigid-as-possible deformation using similarity transformation constraints, while distributing distortions to low-significant regions through an optimization process. The use of object-based significance map and the approach of optimally distributing distortion reduces the need of perfect video segmentation, allowing our method to cope with various cases.

The proposed method is built on previous image retargeting works [5], [6], and the goal of preserving visually salient motions is the same as that of the works [2], [3], [4]. However, our method has substantial differences from these previous methods. First, an object-preserving retargeting scheme is proposed to ease unpleasant motion distortions caused by inconsistent warping on a volumetric object. As shown in Fig. 1, the inconsistent deformation on the moving shuttle (middle figure) is efficiently eased (right figure) by our method. Second, instead of using a pixel-based or grid-based significance map in warping, an object-based one is proposed to preserve content and reduce weaving artifacts. Third, a similarity transformation constraint is adopted to force visually salient objects in 3D spatial-time space to undergo as-rigid-as-possible deformation in warping. Because of these differences, the proposed method has the main contribution of yielding better retargeting results in terms of motion and shape preservation compared with the related methods [4], [7]. The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed approaches.

- *S.-S. Lin, I.-C. Yeh, S.-H. Chang, C.-K. Yeh, and T.-Y. Lee are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC, 701.*
- *C.-H Lin is with Department of Geomatics, National Cheng Kung University, Tainan, Taiwan, ROC, 701.*

Section 4 discusses the experimental results, and Section 5 presents the conclusions, limitation, and future work.

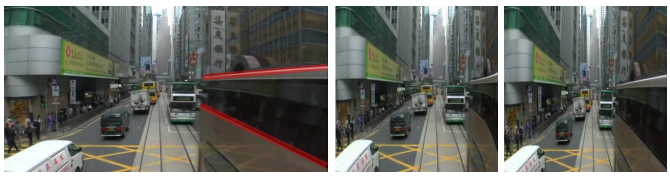


Fig. 1. Motion distortion. Left: the original frame. Middle: the retargeting result generated by Wang et al. [4]. Motion distortion occurred on the moving shuttle due to the inconsistent warping (see the parts marked by red lines on the original frame). Right: the retargeting result with good shape preservation generated by our method.

2 Related Work

Many content-aware retargeting techniques have recently been proposed. We only describe the video retargeting methods in this section. For image retargeting, readers can refer to previous excellent surveys [8], [9]. Content-aware video retargeting methods can be classified into three categories, *cropping*, *seam carving*, and *warping*, based on the operation used in resizing. In the cropping category [10], [11], [12], [13], [14], [15], virtual camera motions, such as panning, zooming, and artificial scene cutting, are produced to maximize the amount of visually salient content within the cropped frames and preserve temporal coherence. However, the methods cannot guarantee that the cutting of meaningful content can be avoided. In the seam carving methods [7], [16], [17], [18], [19], [20], [21], a one-pixel width continuous or discontinuous surface seam with minimal significance in the space-time volume is iteratively carved or inserted to reduce or enlarge the input video to the desired aspect ratio. This technique allows for high flexibility in pixel removal, and thus, can be applied to some interesting applications, such as object removal. However, carving either continuous or discontinuous seams sometimes yields discontinuous artifacts on visually salient objects, causing visual distortion.

In the warping category, methods optimize a mapping or warping using several spatial deformation and temporal coherence constraints to preserve content [2], [3], [4], [22], [23], [24], [25]. In the work of Wolf et al. [22], the original video is resized using non-uniform global mesh warping. Retargeting is formulated as a least-squares problem. The mapping from the source video to the target display is optimized by setting pixel significances which are measured by gradient magnitude as well as detected faces and motions. Therefore, the motions of high-significance objects are preserved, while the low-significance regions are squeezed or stretched. In addition, the position changes of temporally adjacent pixels are penalized in a least-squares optimization to preserve temporal coherence. To efficiently preserve temporal coherence and reduce waving artifacts, Wang et al. [2] proposed to align consecutive frames by estimating inter-frame camera motion and to constrain the relative positions of the aligned frames. They incorporated the motion-aware constraints with an adaptation of the grid mesh warping [5] to preserve visually salient objects. Zhang et al. [23] preserved temporal coherence using a 3D random walk. However, this technique ignores motion information, and may thereby lead to temporal artifacts. Krähenbühl et al. [24] proposed an interactive framework, combining key frame and structure line constraints edited by automatic algorithms, for video analysis and retargeting. Niu et al. [25] consistently resized foreground objects using a motion history map, and maintained background regions by constraining them by the previous frame. This method can well preserve temporal coherence and foreground objects; nevertheless, the preservation of structure lines in the background may be ignored. Lately, Wang et al. [3] combined cropping and warping operators into their framework, where the cropping removes temporally recurring content, and the warping utilized available homogeneous regions to absorb deformations while preserving motions. In addition, without compromising the resizing quality, their later work [4] solved the scalability problem caused by a global optimization over the entire space-time volume. This warping method can yield good retargeting results for many cases. However, an object occupying several grids may suffer from inconsistent scaling and deformation. This may lead to motion distortions, especially for visually salient objects and structure lines. In this study, we ease this spatiotemporal distortion using an object-preserving warping technique.

The concept of object-preserving retargeting has been introduced in the related studies [18], [25], [26], [27]. Tao et al. [26] and Niu et al. [25] extracted and preserved foreground objects in warping based on the fact that users are more interested in the foreground objects. However, preserving only the foreground objects may be unsuitable for images/videos that contain visually salient content in the background. Sun and Ling [27] integrated an objectness map with significance map for the preservation of object completeness in seam carving. The resulting images containing the complete objects can be applied to thumbnail browsing. In Grundmann et al. [18], to preserve objects in video, a spatio-temporal saliency measurement is proposed, in which the input video is segmented into spatio-temporal regions and the saliency is calculated by averaging the frame-based region saliency. Inspired by the work [18], the object-based saliency is used to drive the proposed object-preserving mesh warping for the purpose of consistent object deformation in warping.

3 Content-aware Video Retargeting

Fig. 2 shows the schematic workflow of the proposed method. The basic idea is to resize videos by utilizing the information of object motions. Such information implies inter-frame object correspondence, which

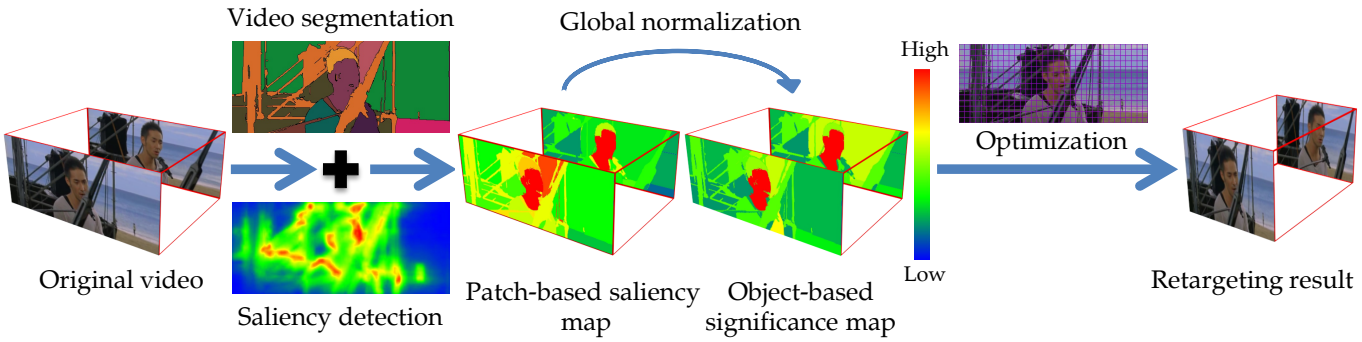


Fig. 2. System workflow. With the aim of consistently preserving content, we utilize an object-based significance map, generated by combining segmented objects and detected saliencies, to force visually salient objects in 3D space-time space to undergo as-rigid-as-possible deformation in video warping. This object-preserving warping can efficiently reduce object inconsistencies in both spatial and temporal deformation.

allows the definition of a significance map for volumetric objects and the consistent preservation of both shapes and motions of volumetric objects in warping. In the proposed method, volumetric objects are first extracted in preprocessing using video segmentation technique. Then, a significance measurement for the segmented objects is performed to generate a significance map for retargeting (Section 3.1). In the significance map generation, the context-aware saliency estimation [28] is adopted to generate a saliency value for each pixel in each frame. Afterward, the significance of each segmented volumetric object is determined by a global normalization process. In the retargeting step, a grid mesh is created to cover the frames, and the proposed object-preserving warping is performed to deform the grids within a volumetric object with high significance as rigidly as possible during the resizing (Section 3.2). We first describe the generation of object-based significance map in Section 3.1, followed by the motion-preserving video warping which is described in Section 3.2.

3.1 Object-based significance map generation

Many techniques of significance map generation for content-aware retargeting have been proposed. Pixels with large gradient magnitudes and saliency values are generally considered as significant pixels [5]. In addition, instead of generating significance map of each frame individually, the pixel significance is measured by considering the content in neighboring frames [2]. However, inconsistent deformation may arise when using such pixel-based (or grid-based) measurement in retargeting. In this study, to consistently preserve content, we adopt an object-based significance measurement, which is inspired by the work [18]. The 3D space-time volume of the original video is initially segmented into several homogenous volumetric objects, each of which is assigned a significance value by a global normalization. With the aid of the globally normalized significance map, volumetric objects can be potentially deformed in a consistent manner.

In the first step, the hierarchical graph-based video segmentation proposed by Grundmann et al. [29] is performed to partition the input video, and the saliency detection approach [28] is adopted to give each pixel an initial significance value. The following is a brief introduction to this video segmentation approach. In [29], the spatiotemporal segmentation begins by over-segmenting a video into volumetric objects using a graph-based segmentation technique. An object graph is then constructed over the initial spatiotemporal segmentation, and a hierarchical tree of segmentation is created by iteratively repeating this process over multiple levels. Segmentation quality is further improved using dense optical flow to guide temporal connections in the object graph. This approach can cope with long videos by including a parallel and out-of-core technique and a clip-based process in the partition scheme. Therefore, the scalability of video segmentation enables us to cope with long videos.

Once the segmented volumetric objects and the pixel salient values are obtained, the object saliency is calculated by simply averaging the salient values of pixels within the volumetric object. Then, each grid is assigned a saliency of object that occupies this grid. This process is called global normalization of object saliency. Figs. 3 and 4 demonstrate the capability of our method to consistently deform objects and reduce weaving, respectively. In Fig. 3, the segmented object occupies grids having similar significance values. Thus, the object can potentially be deformed consistently, and the weaving effects can also be efficiently reduced in warping. Note that generating a perfect video segmentation for all cases is difficult, even when using a state-of-the-art segmentation approach. Fortunately, with the aid of distortion propagation and the object-based significance map, the proposed approach can address the problems caused by unsuccessful object segmentation. The experiment of video retargeting with unsuccessful segmentation will be demonstrated in Section 4. In Fig. 4, we show the advantage of using globally normalized significance map

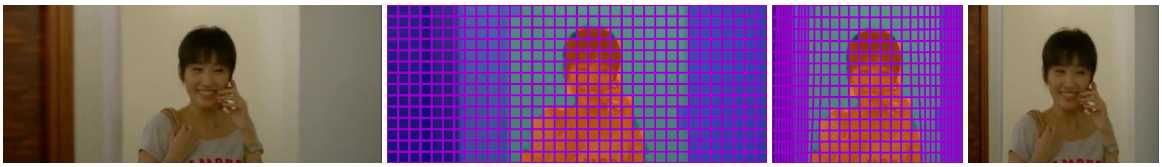


Fig. 3. Consistent deformation. From left to right: original video frame, grid mesh with significance value visualized by color ranging from blue (lowest significance) to red (highest significance), warped grid mesh, and retargeting result.

in retargeting. With this significance map, the weaving artifacts are considerably reduced.

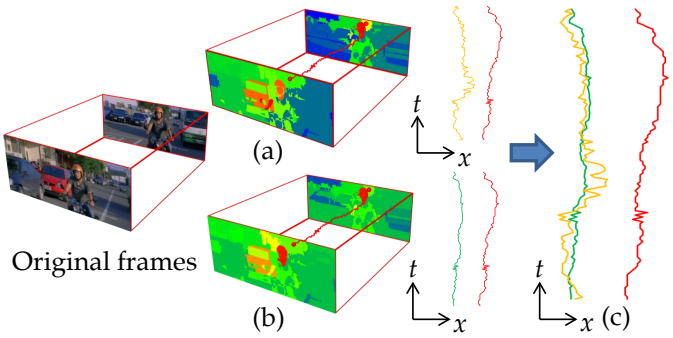


Fig. 4. Comparison of video retargeting using significance map without (a) and with (b) the process of global normalization. The motion trajectories of red point in (a), (b), and the original video are visualized by yellow, green, and red, respectively. These trajectories are shown together in (c) to demonstrate the reduction of weaving artifacts.

3.2 Motion-preserving video warping

An uniform grid mesh $\mathbf{M}^t = (\mathbf{V}^t, \mathbf{E}^t, \mathbf{Q}^t)$ containing a vertex set $\mathbf{V}^t = \{v_1^t, ..., v_{n_v}^t\}$, an edge set $\mathbf{E}^t = \{e_1^t, ..., e_{n_e}^t\}$, and a grid set $\mathbf{Q}^t = \{q_1^t, ..., q_{n_q}^t\}$ is created for t-th frame in the original video, where n_v, n_e, and n_q represent the number of vertices, edges, and grids, respectively. In addition, a set of objects $\mathbf{O}^t = \{object_1^t, ..., object_{n_o}^t\}$ and its corresponding significance values $\mathbf{S}^t = \{S_1^t, ..., S_{n_o}^t\}$ are used in warping, where n_o represents the number of segmented objects. Here, all grid meshes have the same connectivity and are independent of video content. To preserve the spatial content and temporal motions, two energy terms, namely, *spatial content preservation energy* and *temporal coherence preservation energy*, are defined with an optimization solver. These two energy terms are described in the following subsections.

3.2.1 Spatial content preservation energy

Assume that the original video with $m \times n$ resolution is resized into a new video with $m' \times n'$ resolution. The proposed object-preserving warping aims to find a deformed mesh $\mathbf{V}' = \{v_1', ..., v_{n_v}'\}$ for each frame in which the grids in a segmented object are consistently deformed. Three energy terms, namely, *rigid transformation*, *linear scaling*, and *grid orientation*, are defined for this purpose. The term of rigid transformation is used to avoid inconsistent deformation; thus, it is formulated as measuring the rigidity of object in warping:

$$D_{SimT}(\mathbf{M}) = \sum_{i=1}^{n_o}(s_i \times \sum_{\mathbf{e}_j' \in \mathbf{E}(object_i)} \left\|(\mathbf{e}_j' - \mathbf{T}_{ij}\mathbf{C}_i')\right\|^2), \quad (1)$$

where s_i is the significance value of object i. $\mathbf{e}_j'$ and $\mathbf{C}_i'$ represent the deformed edge and the deformed representative edge of object i, respectively. The representative edge is selected as the pivot in object deformation. Generally, the edge closest to the object center is suitable to represent the object and thus suitable to be selected as the representative edge. $\mathbf{T}_{ij}$ is the similarity transformation, containing a scale factor and a rotation factor, between $\mathbf{e}_j$ and $\mathbf{C}_i$. Therefore, this energy measures the changes of edge geometric relations in warping.

To avoid over-deformation on low-significance objects, an energy term with respect to linear scaling is included:

$$D_{LinT}(\mathbf{M}) = \sum_{i=1}^{n_o}((1 - s_i) \times \sum_{\mathbf{e}_j' \in \mathbf{E}(object_i)} \left\|\mathbf{e}_j' - \mathbf{L}\mathbf{T}_{ij}\mathbf{C}_i'\right\|^2), \quad (2)$$

where $\mathbf{L}$ is the matrix of linear scaling $(m \times n) \rightarrow (m' \times n')$. This energy term is defined as measuring the difference between deformed and linear-scaling objects. Thus, this term warps low-significance objects as close as possible to linear scaling. The weighting factor is set to $(1 - s_i)$, and the low-significance object has a large weight to avoid over-deformation. Thus, this term can avoid over-deformations on backgrounds of videos.

The term of grid orientation proposed in [3] is used to avoid skewed artifacts. This term is defined as measuring the grid line bending. Assume that a grid $q : \{v_a, v_b, v_c, v_d\}$ contains two horizontal edges (v_a, v_b), (v_d, v_c) and two vertical edges (v_a, v_d), (v_b, v_c). To measure grid deformation, this term is formulated as the distance of the y component between the vertices of the deformed horizontal edges, and the distance of the x component between the vertices of the deformed vertical edges:

$$\begin{aligned} D_{Ori}(\mathbf{M}) = \sum_{q \in \mathbf{Q}} (&\left\|v_{a_y}' - v_{b_y}'\right\|^2 + \left\|v_{d_y}' - v_{c_y}'\right\|^2 \\ &+ \left\|v_{a_x}' - v_{d_x}'\right\|^2 + \left\|v_{b_x}' - v_{c_x}'\right\|^2), \end{aligned} \quad (3)$$

where the suffixes x and y represent the $x-$ and $y-$component of the vertex position, respectively. The

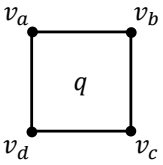
total spatial energy is obtained by summing up the individual spatial energy terms:

$$D_{Sp}(\mathbf{M}) = (\alpha \times D_{SimT}(\mathbf{M}) + (1-\alpha) \times D_{LinT}(\mathbf{M})) + D_{Ori}(\mathbf{M}), \quad (4)$$

where α is the weighting factor for the energy terms D_{SimT} and D_{LinT}. This weighting factor controls how rigid the high-significance objects are. Setting a larger value forces the high-significance objects to be more rigid in warping. To preserve the shapes of visually salient content, α is set to 0.7 in all experiments.

3.2.2 Temporal coherence preservation energy

The temporal energy consists of two energy terms: *warping coherence* and *z-line bending*. The term of warping coherence is formulated as measuring the consistency of object deformation to achieve the goal of consistent warping of volumetric objects:



$$\begin{aligned} &D_{ObjectCoh}(\mathbf{M}) \\ &= \sum_{i=1}^{n_o} \sum_{q \in o_i} (\left\| (v'_{b,t} - v'_{a,t}) - l^{o_i}_{x,init} \right\|^2 \\ &+ \left\| (v'_{c,t} - v'_{d,t}) - l^{o_i}_{x,init} \right\|^2 \\ &+ \left\| (v'_{d,t} - v'_{a,t}) - l^{o_i}_{y,init} \right\|^2 \\ &+ \left\| (v'_{c,t} - v'_{b,t}) - l^{o_i}_{y,init} \right\|^2), \end{aligned} \quad (5)$$

where

$$l^{o_i}_{x,init} = \frac{1}{2 \times n^{o_i}_q} \times \sum_{q \in o_i} ((v'_{b,init} - v'_{a,init}) + (v'_{c,init} - v'_{d,init}))$$

$$l^{o_i}_{y,init} = \frac{1}{2 \times n^{o_i}_q} \times \sum_{q \in o_i} ((v'_{d,init} - v'_{a,init}) + (v'_{c,init} - v'_{b,init}));$$

$l^{o_i}_{x,init}$ and $l^{o_i}_{y,init}$ are the average grid deformations of object o_i in the initial frame (denoted by frame $init$); $n^{o_i}_q$ is the number of grids belonging to object o_i and $v_{a,init}$, $v_{b,init}$, $v_{c,init}$ and $v_{d,init}$ are the corner vertices of grid q in frame $init$. The initial frame of an object is set to its first-appearing frame. Another solution is to search for the optimal frame using a two-pass optimization scheme. In the first pass, each frame is resized individually, and then selects the optimal frame for each object. In the second pass, a global optimization is performed to warp the video. However, this two-pass manner is time-consuming. To consider the algorithm efficiency, we subsequently warp the frames and the first-appearing frame of an object is selected as the initial frame. In Eq. 5, this energy term is defined as measuring the difference of average grid deformations of object o_i in frame t and in the initial frame. Thus, this term can force the volumetric objects, including non-rigid and moving objects, to deform consistently in warping.

Similar to the energy term of grid orientation that measures the grid line bending in spatial space, the energy term of z-line bending measures the grid line bending in temporal space. This term is formulated as the position difference of vertex in frame t and in the neighboring frame $t-1$.

$$D_{LineCoh}(\mathbf{M}) = \sum_{v_i \in \mathbf{V}} \left\| v'_{i,t-1} - v'_{i,t} \right\|^2 \quad (6)$$

The total temporal coherence preservation energy is obtained by summing up these two temporal energy terms:

$$D_{Tep}(\mathbf{M}) = \beta \times D_{ObjectCoh}(\mathbf{M}) + (1-\beta) \times D_{LineCoh}(\mathbf{M}), \quad (7)$$

where β is the weighting factor for these two energy terms. To well preserve the temporal coherence of objects, a large value is assigned to β (β is set to 0.7 in the experiments). Fig. 5 shows the benefit of these temporal constraints through a comparison of retargeting with and without the proposed energy terms. The result shows that the weaving effect is considerably reduced and the temporal coherence is improved by using these energy terms.

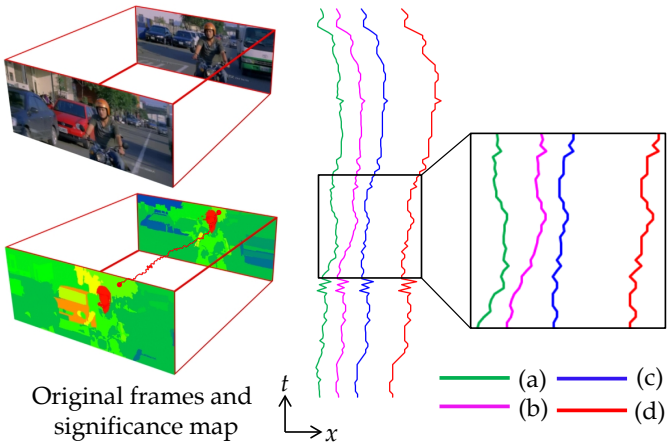


Fig. 5. Comparison of warping optimization. (a) Warping without the temporal energy terms; (b) warping with the z-line bending term; (c) warping with both the z-line bending and warping coherence terms; (d) original motion trajectory.

3.2.3 Minimization of Energy Function

By combining the spatial and temporal energies, the final optimization for frame t is formulated as:

$$\arg\min_{v'_i,t} (D_{Sp} + D_{Tep}), \quad (8)$$

subject to the constraints of the positions of boundary vertices. Here, we assign the same weight to these two energy terms to balance the spatial and temporal energy contribution. In the implementation, similar to [5], we fix the top-left vertex position of the frames and constrain all the boundary vertices of each frame to slide along their respective boundary lines. Finally, a least-squares linear system with a sparse design matrix can be obtained from (8). We solve this system using the conjugate gradient method. The iterative process is terminated

when the movements of boundary and internal vertices are smaller than 0.5 pixels. Since the neighboring frames usually have similar deformations, the result of the previous frame is used as an initial estimation for the next one. Such that the optimization can converge in fewer iterations. In our implementation, the deformed mesh geometry $\mathbf{V}' = \{v'_1, ...v'_{n_v}\}$ is sequentially determined for each video frame, instead of determining all deformed meshes over the entire video volume, making our approach able to cope with long videos. In addition, the error accumulation is very few since only three frames are used in per frame optimization.

4 Experimental Results and Discussion

We tested our algorithm on a desktop PC with Core i5 2.66 GHz CPU and 4 GB memory. For a 688×286 pixel resolution video with 250 frames, the average computation time for video warping is 8.49 seconds (0.034 seconds per frame). Similar to the work [4], the computational complexity of our video warping is $O(N \cdot T)$, where N is the video pixel resolution and T is the number of frames, since the proposed scheme is sequent per-frame resizing.

For a fair comparison, most videos used in the related works were tested in the experiments. Several representative cases that videos contain evident foreground objects and structure lines are shown in Figs. 6, 8, and 10, and the others are attached as accompanying documents. All results were automatically generated using the default parameters, that is, grid resolution is 20 pixles $\times$ 20 pixels and $\alpha = \beta = 0.7$. Please refer the results and comparisons to our accompanying and supplemental videos, especially as the temporal effects are difficult to visualize in still frames.

Fig. 6 provides the results of the proposed retargeting processes, including video segmentation, saliency detection, significance normalization, and mesh warping. Our method resizes videos by utilizing object motions and thereby enabling the consistent preservation of both the shapes and motions of volumetric objects during warping. For example, in Fig. 6, the regions with high significance value (i.e., visually salient content) are well preserved.

Accurate volumetric object extraction is difficult and over-segmentation with unfavorable object boundaries may occur. Fortunately, the proposed warping scheme can ease the difficulty suffering from imperfect video segmentation. For instance, in Fig. 7, the foreground object is partitioned into several patches with unfavorable boundaries. In this case, the effect of inconsistent deformation is still reduced compared with the method using a grid-based significance map. We test videos with over-segmentation and partially incorrect segmentation to further demonstrate the robustness of our approach. The results in Figs. 8 and 9 show that our method does not rely heavily on the accuracy of object segmentation.

Comparison. Most retargeting methods are based on seam carving or mesh warping. Therefore, our method is compared with the standard seam-carving-based method (i.e., improved seam carving (ISC)) [7], and the recent video-warping-based method (i.e., per-frame optimization (PFO)) [4], in addition to the linear scaling (LS). For a fair comparison with the PFO, we used the same resolution for the grid mesh. The comparisons are shown in Fig. 10. The ISC [7] has higher flexibility in pixel removal, and thus, can be applied for object removal. However, the comparisons indicate that the ISC may yield discontinuous artifacts on visually salient objects, sometimes producing noticeable visual distortion. The PFO [4] has the advantage of absorbing distortion by homogeneous regions. However, the human vision is sensitive to inconsistent deformation of visually salient content. In contrast, our method can efficiently eases inconsistent deformation by the object-preserving warping. For examples, in Fig. 10, the roadside in the 1^{st} data,

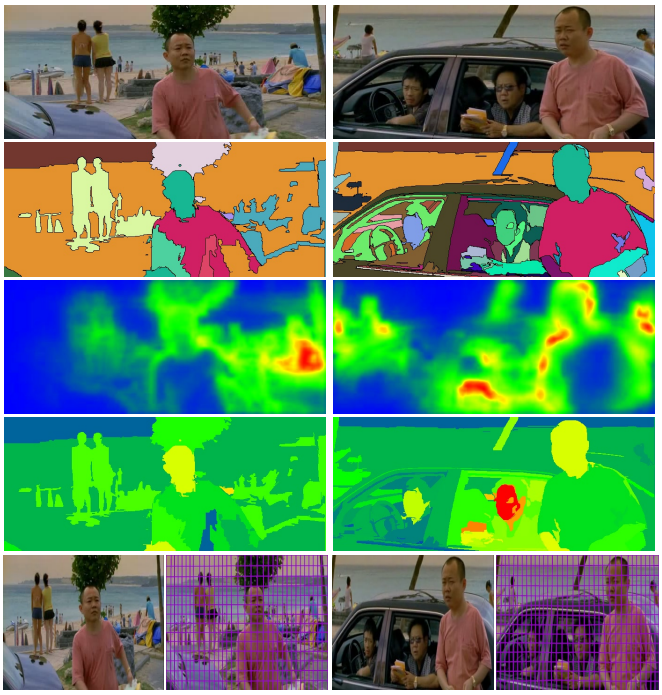


Fig. 6. Retargeting results generated by our approach. From top to bottom: original frames, segmentation results, saliency detection results, significance maps used in retargeting, and our retargeting results.

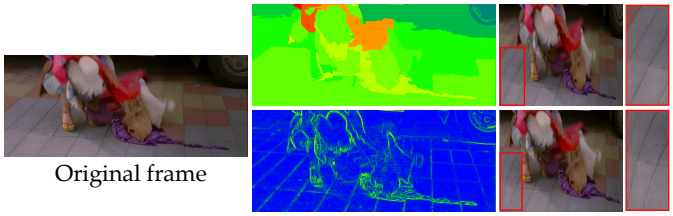


Fig. 7. Comparison of retargeting using object-based (top) and grid-based (bottom) significance maps. Inconsistent deformation occurs in the region marked by red rectangle.

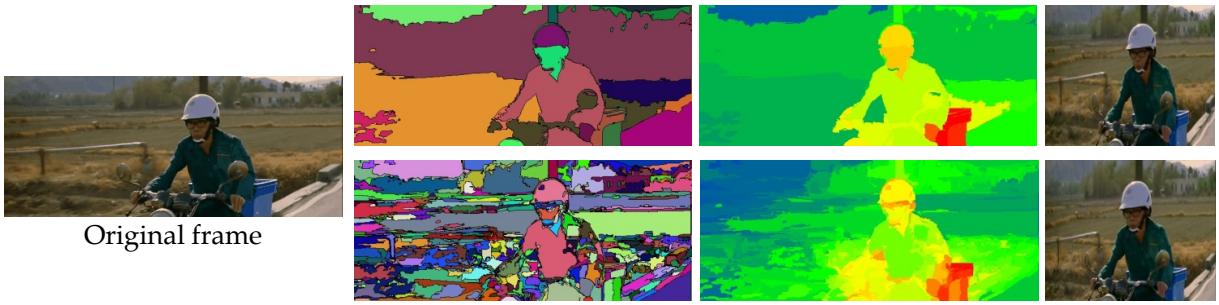


Fig. 8. Video retargeting using general segmentation (top) and over-segmentation (bottom). From left to right: original video frame, segmentation results, significance maps, and retargeting results. The segmented objects are represented by colors.

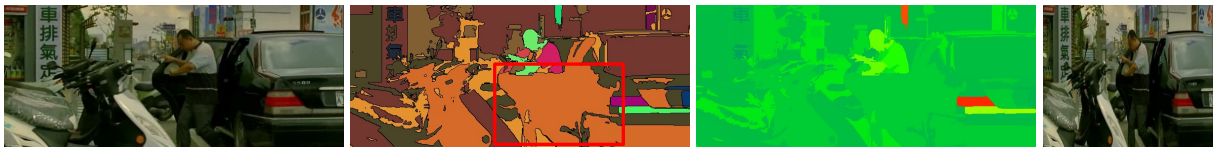


Fig. 9. Video retargeting using unsuccessful object segmentation. From left to right: original video frame, segmentation result, significance map, and our result. The incorrect object segmentation is marked by red rectangle.

the moving shuttle in the 2^{nd} data, the red box and paper in the 3^{rd} data, the white lines in the 4^{th} data, and the shape of foreground object in the 5^{th} data are well preserved. Moreover, using temporal coherence constraints with normalized significance map greatly reduces waving artifacts (refer to our supplemental videos). These properties enable our method to generate better results compared with those from related methods.

The goal of preserving motions in warping is the same as the work [4], therefore a quantitative analysis on motion preservation is conducted by using correlation coefficient. The correlation coefficient between two sets of motion trajectories m_i and m_j is defined as $Corr(m_i, m_j) = \frac{Cov(m_i, m_j)}{\sigma_{m_i}\sigma_{m_j}}$, where $Cov(m_i, m_j)$ means covariance between m_i and m_j, and σ_m represents the standard deviation of m. In this experiment, two clips containing evident foreground objects and structure lines are tested. Several feature points in the original clip and the retargeting results generated by our method and PFO are manually selected. The correlation coefficient between the motion trajectories formed by the selected feature points in the generated retargeting results and original clip are calculated. The analyses shown in Fig. 11 indicate that our results are closer to the original clips compared with the results of PFO. To further show how important the proposed spatial and temporal energy terms and object-based significance map are in shape and motion preservation, we compare with the approach that use PFO warping and the proposed significance map. The results show that using our significance map in PFO (Fig. 12, bottom figure) can improve PFO's retargeting quality (Fig. 12, middle figure). However, shape and motion distortions still occur (see the motion shuttle). The distortions can be eased by using the proposed energy terms in warping (Fig. 12, top figure). This experiment shows that not only the significance map but the energy terms are contributions to better shape and motion preservation.

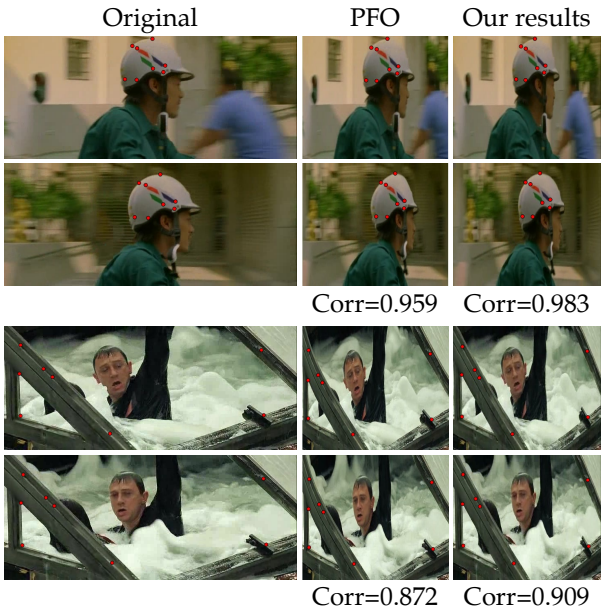


Fig. 11. Quantitative analysis using correlation coefficient (denoted by Corr). The correlation between the trajectories of the selected points (marked by red) in the original clip and generated retargeting clip is calculated.

User Study. A user study involving 90 participants, aged 20 to 47 years old, was conducted to evaluate our method. We used the survey system and the paired comparison method provided by Rubinstein et al. [9].

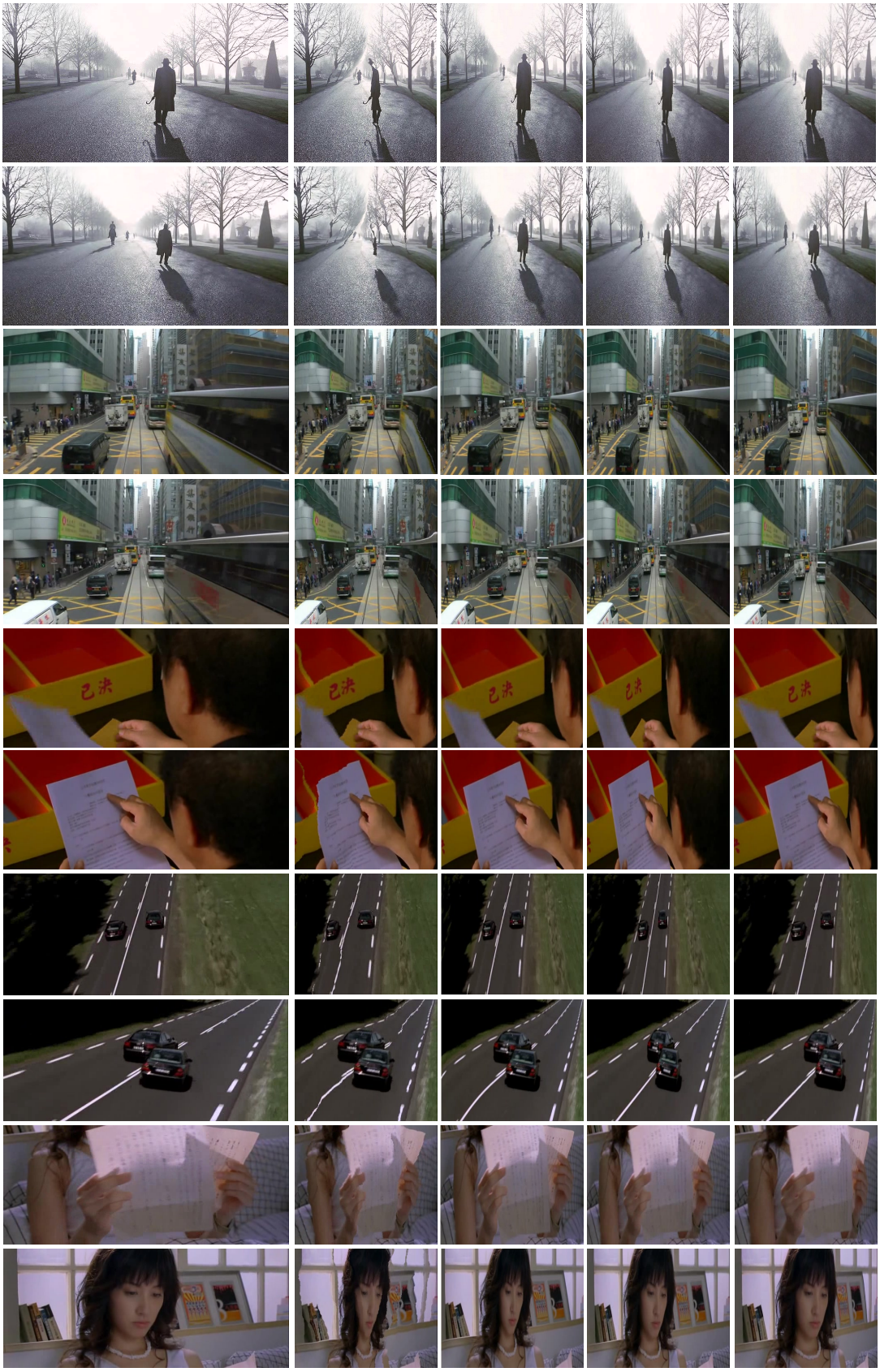


Fig. 10. Comparisons with the related methods, including improved seam carving (ISC) [7], per-frame optimization (PFO) [4], and linear scaling (LS).

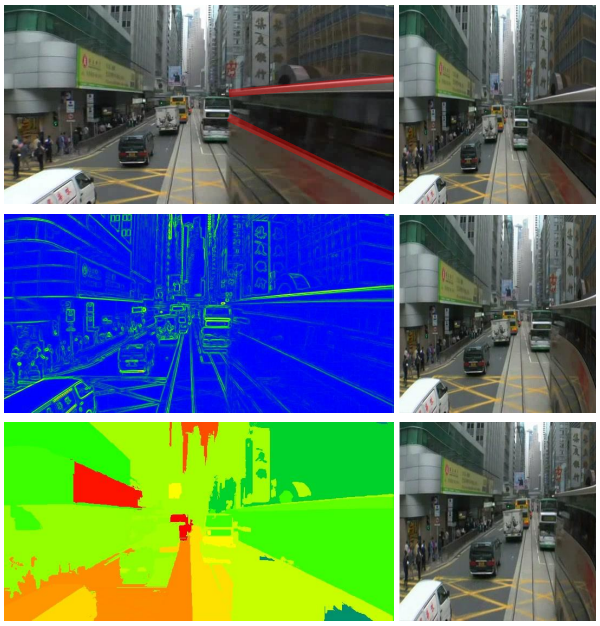


Fig. 12. Comparison of video retargeting using different significance maps and warping approaches. Top: Original video frame and our retargeting result; middle: significance map used in [4] and the retargeting result using PFO; bottom: our significance map and the retargeting result using PFO.

Participants are shown two retargeted videos side by side at one time, and asked to choose the one they like better. Following [9], we use videos having the attributes that can be mapped to the content-aware retargeting objectives: preserving content and preserving structure. It is likely difficult to use a dataset containing too many attributes and video categories in the user study. Participants are likely to loose their patience/concentration in a long user study. Therefore, the video dataset is made up of only eight videos having the main attributes, *evident foreground objects* and *structure lines* (see the supplemental video). From the number of votes shown in Fig. 13, this survey indicates that our results are better than those generated by the ISC, PFO, and LS for the videos having structure lines and evident foreground objects.

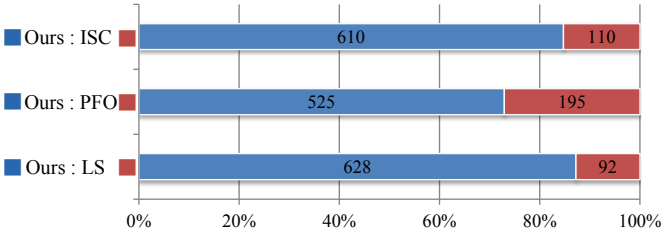


Fig. 13. The total number of votes for our method and the methods ISC [7], PFO [4], and LS.

5 Conclusions, Limitation and Future Work

A novel object-preserving warping for content-aware video retargeting is presented. In the optimization of grid mesh warping, the spatial content preservation constraints force the visually salient content to undergo as-rigid-as-possible deformation, and the temporal coherence preservation constraints considerably reduce weaving artifacts. Moreover, the optimization process with the normalized significance map propagates distortions to low-significant regions. These processes significantly ease the problems of unpleasant motion deformations caused by inconsistent warping, enabling our approach to effectively cope with videos containing dense information and structure lines. The comparisons and user study clearly show the superiority of our method over the related methods in terms of content preservation. At present, our approach may have the problem of over-constraining when the input video fills with global structure lines, as shown in Fig. 14. In such a case, the result is similar to linear rescaling, because all the structure lines and most objects are deformed rigidly. As for video segmentation, we have demonstrated that the problems caused by imperfect object segmentation can be handled by our warping scheme, meaning that the retargeting quality does not rely heavily on the accuracy of object segmentation. In the future, we plan to extend our retargeting scheme to stereo videos and multi-temporal geospatial data.

References

[1] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Comput. Graph. Forum*, vol. 28, no. 7, pp. 1897–1906, 2009.

[2] 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, 2009.

[3] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 90:1–90:9, 2010.

[4] Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee, "Scalable and coherent video resizing with per-frame optimization," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 88:1–88:8, 2011.

[5] 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, 2008.

[6] S.-S. Lin, I.-C. Yeh, C.-H. Lin, and T.-Y. Lee, "Patch-based image warping for content-aware retargeting," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 359–368, 2013.

[7] 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, 2008.

[8] A. Shamir and O. Sorkine, "Visual media retargeting," in *ACM SIGGRAPH ASIA 2009 Courses*, 2009, pp. 11:1–11:13.

[9] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 160:1–160:10, 2010.

[10] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," in *Proceedings of the 14th annual ACM international conference on Multimedia*, 2006, pp. 241–250.

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

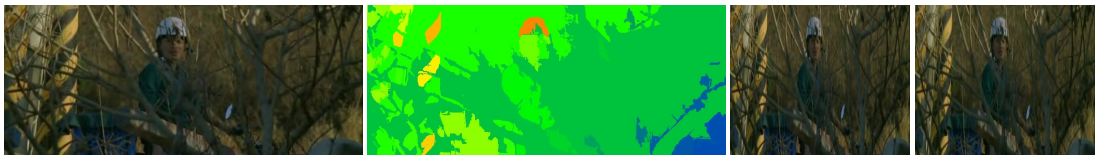


Fig. 14. Retargeting result for the video that fills with global structure elements. From left to right: original video frame, significance map, linear rescaling, and our result.

[12] T. Lu, Z. Yuan, Y. Huang, D. Wu, and H. Yu, "Video retargeting with nonlinear spatial-temporal saliency fusion," in *Proceedings of IEEE International Conference on Image Processing*, 2010, pp. 1801–1804.

[13] Y. Luo, J. Yuan, P. Xue, and Q. Tian, "Salient region detection and its application to video retargeting," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2011, pp. 1–6.

[14] M. Frankovich and A. Wong, "Enhanced seam carving via integration of energy gradient functionals," *Signal Processing Letters*, vol. 18, no. 6, pp. 375–378, 2011.

[15] S. Kopf, T. Haenselmann, J. Kiess, B. Guthier, and W. Effelsberg, "Algorithms for video retargeting," *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 819–861, 2011.

[16] A. Shamir and S. Avidan, "Seam carving for media retargeting," *Commun. ACM*, vol. 52, no. 1, pp. 77–85, 2009.

[17] 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. Cir. and Sys. for Video Technol*, vol. 19, no. 11, pp. 1588–1597, 2009.

[18] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Discontinuous seam-carving for video retargeting," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 569–576.

[19] Y. Hu and D. Rajan, "Hybrid shift map for video retargeting," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 577–584.

[20] H.-M. Nam, K.-Y. Byun, J.-Y. Jeong, and K.-S. Choi, "Low complexity content-aware video retargeting for mobile devices," *IEEE Trans. Consumer Electronics*, vol. 56, no. 1, pp. 182–189, 2010.

[21] W.-L. Chao, H.-H. Su, S.-Y. Chien, W. H. Hsu, and J.-J. Ding, "Coarse-to-fine temporal optimization for video retargeting based on seam carving," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2011, pp. 1–6.

[22] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proceedings of the Eleventh IEEE International Conference on Computer Vision*, 2007, pp. 1–6.

[23] 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.

[24] 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, 2009.

[25] Y. Niu, F. Liu, X. Li, and M. Gleicher, "Warp propagation for video resizing." in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 537–544.

[26] C. Tao, J. Jia, and H. Sun, "Active window oriented dynamic video retargeting," in *In Proceedings of the Workshop on Dynamical Vision*, 2007.

[27] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 1511–1518.

[28] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 2376–2383.

[29] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 2141–2148.