Spatiotemporal Grid Flow for Video Retargeting

Bing Li, Ling-Yu Duan, *Member, IEEE*, Jinqiao Wang, *Member, IEEE*, Rongrong Ji, *Member, IEEE*, Chia-Wen Lin, *Senior Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*— Video retargeting is a useful technique to adapt a video to a desired display resolution. It aims to preserve the information contained in the original video and the shapes of salient objects while maintaining the temporal coherence of contents in the video. Existing video retargeting schemes achieve temporal coherence via constraining each region/pixel to be deformed consistently with its corresponding region/pixel in neighboring frames. However, these methods often distort the shapes of salient objects, since they do not ensure the content consistency for regions/pixels constrained to be coherently deformed along time axis. In this paper, we propose a video retargeting scheme to simultaneously meet the two requirements. Our method first segments a video clip into spatiotemporal grids called grid flows, where the consistency of the content associated with a grid flow is maintained while retargeting the grid flow. After that, due to the coarse granularity of grid, there still may exist content inconsistency in some grid flows. We exploit the temporal redundancy in a grid flow to avoid that the grids with inconsistent content be incorrectly constrained to be coherently deformed. In particular, we use grid flows to select a set of key-frames which summarize a video clip, and resize subgrid-flows in these key-frames. We then resize the remaining nonkey-frames by simply interpolating their grid contents from the two nearest retargeted key-frames. With the key-frame-based scheme, we only need to solve a small-scale quadratic programming problem to resize subgrid-flows and perform grid interpolation, leading to low computation and memory costs. The experimental results demonstrate the superior performance of our scheme.**

***Index Terms*— Video retargeting, video warping, dynamic programming, quadratic programming.**

I. Introduction

VIDEO retargeting aims to resize a video to a desired resolution or aspect ratio. With proliferation of accessing videos on various devices with different resolutions and aspect ratios (e.g. PCs, TVs, Tablets and mobile phones), video retargeting becomes an important tool to adapt a video to the displays of these devices. It is also driven by the need of convenient human-device interactions. For example, a viewer might frequently change the size of display window, leading to changes on resolution and aspect ratio of video. However, existing video scaling schemes such as uniform scaling or cropping often distort or discard salient object in practice. To this end, much research effort [1]–[30] has been devoted to content-aware retargeting, which non-uniformly deforms a video/image, such that the shapes of salient objects are preserved at the cost of distorting unimportant content.

The simplest approach to content-aware video retargeting is to independently resize individual frames of a video in a content-aware manner [16]. However, this inevitably introduces temporal incoherence artifacts such as jittering, since even small frame to frame changes incur noticeable temporal discontinuity. To address this issue, the methods proposed in [17]–[24] sequentially resize frames, where each frame is constrained to be deformed similarly to several neighboring frames to avoid jittering artifacts. However, since they only utilize local temporal information in a time window to ensure temporal coherence, they would make the same objects in adjacent frames undergo visible incoherent deformation in and beyond the time window, when the window is shorter than the duration of camera/object motions.

In contrast, the methods proposed in [25]–[29] use information extracted from the entire video clip to ensure temporal coherence. Notably, recent works in [27]–[29] first independently divide each frame into grids and align the grids between consecutive frames via image registration. Consequently, through the propagation of registration between consecutive frames, each grid is automatically aligned to the corresponding ones in other frames. Based on the alignments, the methods formulate an optimization over all frames, which constrains the aligned grids to be deformed coherently across the entire video. Thanks to such a global manner, these methods can achieve temporal coherence. The global manner, however, often degrades the performance of preserving salient objects for videos with significant motions, compared to the results of independently resizing.

To keep good shape preservation results of independently resized frames, regions/pixels, which are constrained to be coherently deformed along the time axis, should exhibit *consistency in content*. However, existing methods usually fail to do so due to two limitations. First, they may wrongly align regions/pixels containing inconsistent content between consecutive frames. Although the works in [27]–[29] perform motion estimation to align grids, their grid partitions ignore the correlations between frames and thus lead to many inaccurate

Manuscript received June 26, 2013; revised November 23, 2013; accepted January 28, 2014. Date of publication February 11, 2014; date of current version February 25, 2014. This work was supported in part by the Chinese Natural Science Foundation under Grants 61271311, 61121002, 61210005, and 61273034, and in part by Supervisor Award Funding for Excellent Doctoral Dissertation of Beijing (20128000103). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marios S. Pattichis. *(Corresponding author: L.-Y. Duan.)*

B. Li is with Institute of Computing Technology, the University of Chinese Academy of Sciences, Beijing 100190, China, and Institute of Digital Media, School of Electronic and Engineering Computer Sciences, Peking University, Beijing 100871, China (e-mail: bing.li@pku.edu.cn).

L.-Y. Duan, R. Ji, and W. Gao are with the Institute of Digital Media, School of Electrical and Engineering Computer Sciences, Peking University, Beijing 100871, China (e-mail: lingyu@pku.edu.cn; rrji@pku.edu.cn; wgao@pku.edu.cn).

J. Wang is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Bejing 100190, China (e-mail: jqwang@nlpr.ia.ac.cn).

C.-W. Lin is with the 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.2014.2305843

temporal alignments. That is, an object appearing in a grid may be only partially covered by the grid's aligned grids in adjacent frames. Second, in the case of such inaccurate temporal alignments, using all frames to guide the deformation of pixel/region is inappropriate, since this will largely increase the possibility that a region/pixel is improperly enforced to be deformed coherently with ones containing inconsistent content.

Overall, to maintain temporal coherence without degrading the performance on shape preservation of salient objects, it is necessary for the retargeting methods to explicitly maximize the content consistency of regions/pixels constrained to be deformed coherently across a video. This was, however, rarely considered in the existing methods to the best of our knowledge. One effective approach is to segment a video into a set of spatio-temporal regions, each of which representing an object. Then the spatio-temporal regions are resized and recomposed to adapt the video to a desired size. However, this needs accurate video object segmentation which is computationally expensive and technically challenging.

In this paper, we propose a novel video retargeting method to efficiently maintain temporal coherence without degrading shape preservation of salient objects. The contribution of the proposed method is three-fold. First, we propose to divide a video into spatio-temporal grids, namely, *grid flows* to well represent spatio-temporal regions of visual contents, such that content consistency among grids for each grid flow is maximized during grid partitioning. Second, owing to the high redundancy in temporal information, we propose a key-frame selection scheme based on the Bayesian networks of grid flows, such that the optimization of retargeting is performed over a small set of key-frames, rather than over the entire video. This does not only suppress the negative impact of content inconsistency, but also greatly reduces memory and computation costs. Third, we formulate the grid-flow-guided retargeting for the selected set of key-frames as a convex quadratic programming problem, and propose an efficient iterative solver. We also propose to resize the remaining non-key-frames by using simple grid interpolation to achieve fast and high-performance retargeting.

The rest of this paper is organized as follows. Sec. II surveys the recent image and video retargeting schemes. The problem formulation for the proposed framework is described in Sec. III. Sec. IV elaborates the grid flow construction scheme. Sec. V presents the proposed grid-flow-guided retargeting scheme. Sec. VI reports the experimental results. Finally, conclusions are drawn in Sec. VII.

II. Related Work

Recently many content-aware image and video retargeting methods have been proposed. Since most existing video retargeting methods are extensions of image retargeting algorithms, we will briefly review the image retargeting methods first.

A. Image Retargeting

Generally speaking, there exist quite a few content-aware image retargeting methods such as cropping, seam carving and warping. Among these methods, cropping-based methods [1]–[4] often adopt attention models to detect relatively important image regions and then only crop the most important region to display. In contrast, seam carving [5], [6] aims to carve a group of seams with the lowest energy values iteratively from an image based on the energy map derived from the image. In the warping-based methods [7]–[13], an image is first segmented into pixels or regions (e.g., grids), which are then non-uniformly deformed through an optimization process, in which spatial constraints are imposed to minimize the deformation distortion on important pixels/regions. In many cases, these cropping, seam carving and warping methods can achieve promising performance. However, they may fail when encountering unexpected image contents. For example, seam carving may distort a structural object while downsizing an image, while cropping may fail when processing an image containing several salient targets.

To address these problems, some hybrid methods put efforts on combining several retargeting methods to overcome the drawbacks of individual retargeting methods. For example, Rubinstein et al. [14] proposed to combine different retargeting methods including uniform scaling, cropping, and seam craving to improve retargeting quality. Dong et al. [15] integrated seam craving with scaling to preserve the structural objects. In [31], seam carving was combined with warping for thumbnails browsing.

B. Video Retargeting

Compared with image retargeting, video retargeting is much more challenging due to the additional temporal dimension. In video retargeting, video can be represented as a spatio-temporal 3D volume while various 2D frames are tightly correlated with each other as the same object may appear in several neighboring frames. If each frame is retargeted independently, an object may be incoherently deformed in different frames, leading to serious temporal incoherence such as waving and flickering artifacts which are visually annoying. Therefore, it is necessary to maintain temporal coherence by exploiting the temporal information.

Generally speaking, existing video retargeting methods can be roughly grouped into two categories according to their methodologies of using the temporal information. "Local" methods only use temporal information from a few neighboring frames, whereas "global" methods usually utilize the temporal information from the entire video clip. For the local methods, some methods [17]–[23] resize each frame sequentially by using the cues from its previous or next frame. For example, Zhang et al. [20] and Wolf et al. [17] proposed pixel-based warping methods for video retargeting. When warping a frame, its previous frame is used to build temporal constraints, which constrains the pixels that lie in the same locations in these two neighboring frames to undergo similar location changes. In [22], image-based seam carving [5] was extended by using the carved seams of each frame to modify the energy map of its next frame. That is, areas that are similar to the carved seams are assigned with lower energy while the others are assigned with higher energy. In this way, the same object can be carved coherently in consecutive frames. In [32]

and [33], image cropping is applied to video by smoothing the trajectory of the cropped window in adjacent frames. However, the retargeting results heavily rely on the content of the first frame and the salient objects in the following frames are often distorted.

Instead of using only the information from the previous or next frame, some methods proposed to utilize several neighboring frames to achieve better performance. For instance, Krähenbühl *et al.* [18] and Greisen *et al.* [19] improved the method in [17] by estimating the pixel importance through averaging the saliency maps of several adjacent frames. In [21], the pixel importance was calculated in a similar manner and the traditional image warping algorithm was extended by coherently deforming the video content at the same locations between consecutive frames. Nevertheless, these methods usually fail to process a video with large motions when the number of used neighboring frames is too short to cover the duration of camera/object motion.

To sum up, the local methods can achieve high computing efficiency since they only require a few frames at a time. However, a main drawback of these methods lies in the fact that the temporal information in a local window may be insufficient to ensure satisfactory temporal coherence in the retargeted video.

To overcome this problem, the global methods propose to use the temporal information from the entire video to extend image retargeting methods such as cropping, seam carving and warping to video. For example, global video cropping methods [34], [35] extended image cropping to video by smoothing the trajectory of cropped windows over the entire video. However, the common drawback of video cropping methods is that important video contents (e.g. salient objects) may be cropped out and false camera motions may be introduced.

Rubinstein et al. [25] extended seam carving [5] to video by designing temporal constraints over the entire video clip. In terms of the temporal constraints, the simultaneously carved seams were required to be continuous between consecutive frames, so as to form a spatial-temporal surface across the video. However, such temporal constraints often cause temporal artifacts and distortion of salient objects due to inaccurate temporal alignments.

In [26]–[29], image-based warping methods were extended to warp a video by maintaining consistent deformation of aligned pixels/grids across the entire video clip. Moreover, motion estimation is performed to align pixels/grids which contain the same video content. For example, Yen et al. [26] employed feature-based camera motion estimation to align pixels across a video shot. Then, an image-based warping method was extended to warp a video at pixel-level, i.e., non-uniformly scaling pixels in each frame while constraining the aligned pixels to be scaled coherently. Different from [26], all frames are divided by predefined grids in [27]–[29]. In [28] and [27], motion estimation (e.g., camera motion estimation) is also performed to align grids between two consecutive frames. Then, an optimization was formulated over all grids of the video to simultaneously warp all frames, where deformation differences of aligned grids are penalized between each pair of consecutive frames. By propagating temporal constraints through consecutive frames, the deformation of each grid would be influenced by all frames in the global optimization. In contrast to [27], [28], the method proposed in [29] utilizes flow estimation to aligned grids across the entire video, and then directly constrains the aligned grids to undergo consistent deformation.

Although the above-mentioned global warping methods can achieve temporal coherence for most videos, a common problem of these methods is that they usually may not achieve satisfactory shape preservation of salient objects because they may wrongly enforce different contents to be deformed coherently between frames. Besides, their performances are sensitive to pixel/region alignment inaccuracy. To better preserve the shapes of salient objects, cropping was further combined with warping in [28] and [29]. Yet, the problem of cropping such as creating false camera motion is also introduced. In addition, most global warping methods are computationally expensive, since these methods need to solve optimal retargeting of individual pixels/regions for the entire video.

In this work, we propose a novel video warping method, which is different from existing methods in two aspects. First, unlike the traditional two-step methods that first independently segment each frame into regions/pixels, and then align regions/pixels between frames, our method segments a video into spatio-temporal grid flows, for which content consistency between grids is maximized. Second, instead of using local or global temporal information, we use information from selected key-frames which summarize information of a video to maintain the temporal coherence of retargeted video. In this way, the proposed key-frame-based retargeting scheme significantly reduces the computational complexity while successfully maintaining temporal coherence and shape preservation.

III. Problem Formulation

Let $\mathbf{I} = \{I^t\}_{t=1}^{N_T}$ denote a video, where I^t is the t-th frame and N_T is the frame number. Our goal is to maintain temporal coherence without degrading shape preservation of salient objects when retargeting $\mathbf{I}$ to a desired size. We formulate this task as the following optimization problem:

$$\min \sum_{t=1}^{N_T} D_S^t + \sum_{t=1}^{N_T} \sum_{t' \neq t}^{N_T} D_T^{t,t'}, \tag{1}$$

where D_S^t is a spatial distortion function measuring the shape deformation distortion of salient objects in I^t, and $D_T^{t,t'}$ is a temporal distortion function measuring the temporal incoherence between retargeted frames I^t and $I^{t'}$.

The challenge of solving (1) lies in that minimizing $D_T^{t,t'}$ may negatively influence the minimization of D_S^t. In particular, to minimize $D_T^{t,t'}$, we need to align regions/pixels between frames and constrain them to be deformed coherently. Consequently, the scaling of each frame (i.e., D_S^t) is not only guided by the importance of its regions/pixels, but also influenced by the aligned ones (i.e., the correspondences) in other frames. However, the scaling of a pixel/region may contradict the guidance of its importance, when its corresponding pixels/regions

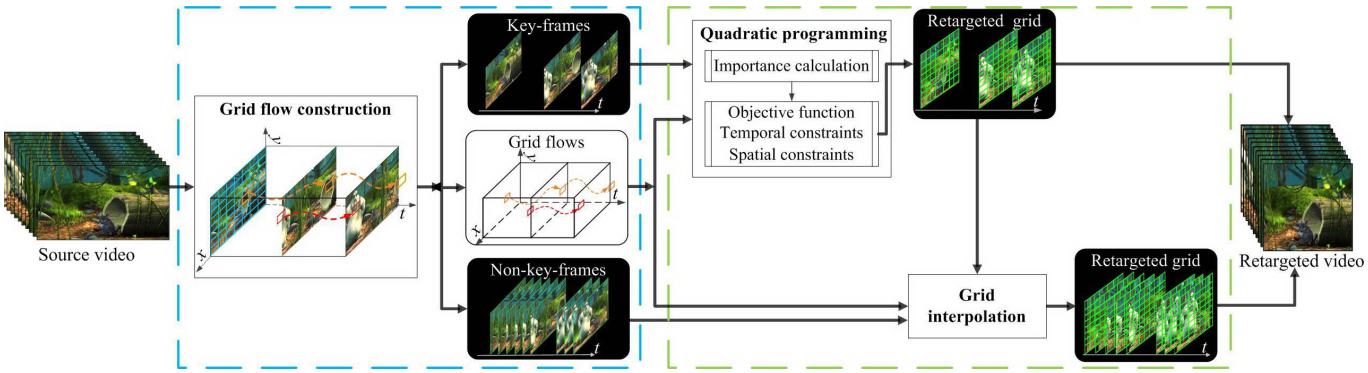


Fig. 1. The framework of the spatiotemporal grid flow for video retargeting.

in neighboring frames contain inconsistent contents. Such contradiction would cause shape deformation distortion of salient objects.

To avoid degrading shape preservation, we should ensure the aligned regions/pixels contain consistent contents. In addition, since pixel-wise alignments are computationally expensive, we propose to perform alignments at the grid level rather than at the pixel level. To this end, as illustrated Fig. 1, we divide a video into spatio-temporal grid flows so that content consistency is maintained among the grids in each grid flow. To construct such grid flows, we partition grids for each frame according to the contents of grids in the previous frame. A set of grids covering consistent contents between frames are aligned to form a grid flow. Under such a circumstance, new grid flows are created if new content appears in a frame.

With the proposed grid flows, we can resize a video like a global method. That is, we formulate an optimization over all frames, where grids in each frame are non-uniformly deformed to preserve salient contents, whereas grids in each grid flow are coherently deformed to ensure temporal coherence. Let $\mathbf{g}_i = \{g_i^t\}_{t=t_i^s}^{t_i^e}$ denote the i-th grid flow, where g_i^t is the grid in I^t, t_i^s and t_i^e respectively denote the frame indices of the initial and final frames containing $\mathbf{g}_i$. We reformulate (1) as

$$\min \sum_{t=1}^{N_T} \sum_i s_i^t \cdot D_S'(g_i^t, \tilde{g}_i^t) + \sum_{t=1}^{N_T} \sum_{t' \neq t} \sum_i D_T'(\tilde{g}_i^t, \tilde{g}_i^{t'}), \quad (2)$$

where $\tilde{g}_i^t$ is the deformed version of g_i^t in the retargeted video, $D_S'(g_i^t, \tilde{g}_i^t)$ a spatial distortion function measuring the deformation discrepancy between g_i^t and $\tilde{g}_i^t$, and $D_T'(\tilde{g}_i^t, \tilde{g}_i^{t'})$ the temporal distortion function measuring the deformation discrepancy between $\tilde{g}_i^t$ and $\tilde{g}_i^{t'}$, s_i^t the weight (i.e., the importance) of grid g_i^t.

However, due to the coarse granularity of grids, there still may exist content inconsistency in some grid flows. Using such global temporal information to ensure temporal coherence may negatively influence shape preservation of salient objects, since it increases the possibility that regions/pixels containing inconsistent content are constrained to be deformed coherently between frames. To address the problem, we exploit temporal redundancy. That is, according to grid flows we select a set of key-frames which summarize the major information of video, and then reformulate the optimization problem in (2) over the set of key-frames as

$$\min \sum_{t \in \mathbf{R}} \sum_i s_i^t \cdot D_S'(g_i^t, \tilde{g}_i^t) + \sum_{t \in \mathbf{R}} \sum_{t' \neq t} \sum_i D_T'(\tilde{g}_i^t, \tilde{g}_i^{t'}), \quad (3)$$

where $\mathbf{R}$ represent the key-frames.

The rest of non-key-frames are subsequently resized via simple grid interpolation based on the retargeted results of their two nearest key-frames at low computational complexity.

IV. Grid Flows Construction

We construct the grid flows of a video based on Bayesian networks. We accordingly formulate the grid flow construction as a problem of finding the longest path in a directed graph, which can be efficiently solved via dynamic programming. The details are elaborated below.

A. Formulation of Grid Flow Construction

Suppose the first frame I^1 is divided into n grids $g_1^1, \ldots, g_n^1$. We initialize grid flows as $\mathbf{g}_i = \{g_i^1\}$. Then we sequentially determine grid g_i^{t+1} for I^{t+1} according to the content covered by g_i^t, and update grid flows $\mathbf{g}_i = \mathbf{g}_i \bigcup g_i^{t+1}$ by aligning the grids. Sequentially, we construct new grid flows $\mathbf{g}_n, \mathbf{g}_{n+1}, \ldots$, once new content appears in I^{t+1}.

Suppose $\mathbf{G} = \{\mathbf{g}_i\}_{i=1}^N$ represents the complete set of grid flows in a video, where N is the number of grid flows. We assume that g_i^{t+1} depends on I^t, I^{t+1} and g_i^t only, and grid flows are mutually independent. Based on the assumption, we model the process of constructing grid flows by a Bayesian network as shown in Fig. 3. To specify the Bayesian network, we need to formulate the construction of a grid flow, and determine when to terminate an existing grid flow and when to initiate a new one.

Constructing a grid flow: As shown in Fig. 2, we determine g_i^{t+1} by finding out the rectangular region in I^{t+1} containing the most consistent content as g_i^t. Thus, searching for g_i^{t+1} is to estimate the displacement of g_i^t from I^t and I^{t+1}:

$$P(g_i^{t+1} | g_i^t, I^t, I^{t+1}) \quad s.t. \quad g_i^{t+1} = \Psi(g_i^t, \mathbf{o}), \quad (4)$$

where $P(g_i^{t+1} | g_i^t, I^t, I^{t+1})$ is the posterior probability of g_i^{t+1} given I^{t+1}, I^t and g_i^t, $\mathbf{o}$ stands for the 2D spatial displacement

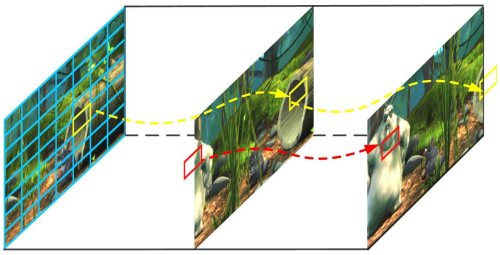


Fig. 2. Illustration of constructing grid flows, where two grid flows are indicated in different colors. We uniformly divide the first frame into grids. Then we track each grid and generate grids for the successive frames according to grid motion. New grid flows are created if new content appears (See the red grid flow).

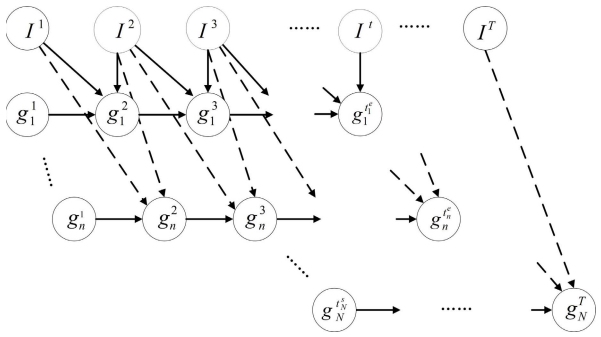


Fig. 3. Bayesian network for grid flow construction, where nodes represent variables (e.g. grid g_i^t, frame I^t), both dashed arrows and solid ones represent conditional dependencies between these variables.

vector, $\Psi(g_i^t, \mathbf{o})$ represents the grid obtained by translating g_i^t with $\mathbf{o}$.[1]

For $g_i^{t'} = \Psi(g_i^t, \mathbf{o})$, we measure $P(g_i^{t'}|I^t, I^{t'}, g_i^t)$ which is the visual content similarity between grid g_i^t and $g_i^{t'}$. More specifically, we represent the content in a grid by using color histogram. We then employ histogram intersection to measure the visual similarity between grids. Let $\mathbf{H}_i^t = [H_i^t(1), \ldots, H_i^t(d)]$ denote the color histogram of g_i^t, $P(g_i^{t'}|I^t, I^{t'}, g_i^t)$ is calculated by:

$$P(g_i^{t'}|I^t, I^{t'}, g_i^t) = \sum_{j=1}^{d} \frac{\min(H_i^{t'}(j), H_i^t(j))}{\| \mathbf{H}_i^t \|_1}, \tag{5}$$

where L1-norm $\| \mathbf{H}_i^t \|_1$ counts the number of pixels in g_i^t, and the number of color bins d is set to 80 empirically.

Terminating a grid flow: When the content represented by $\mathbf{g}_i$ completely disappears in I^{t+1}, the evolution of g_i^t in I^{t+1} does not exist. In this case, we shall terminate $\mathbf{g}_i$, if

$$P\left(g_i^{t+1} = \Psi(g_i^t, \mathbf{o})|g_i^t, I^t, I^{t+1}\right) < \varepsilon \quad \forall \mathbf{o}, \tag{6}$$

where ε is a small positive value.

Initiating a new grid flow: If new content appears in I^{t+1}, a new grid flow should be initiated to represent and track the new content. Suppose g_{new} contains new content in I^{t+1}, any grid in I^t would yield small $P(g_i^{t+1} = g_{new}|g_i^t, I^t, I^{t+1})$. Therefore, we initiate a new grid flow with the prior probability as follows:

$$P(g_{n'+1}^{t+1}) = 1 - \frac{1}{n'} \sum_{i=1}^{n'} P(g_i^{t+1} = g_{new}|g_i^t, I^t, I^{t+1}), \tag{7}$$

where n' is the number of grid flows after I^{t+1} is processed, $g_{n'+1}^{t+1}$ the initial grid of new grid flow $\mathbf{g}_{n'+1}$.

Objective function: We formulate the Bayesian network (See Fig. 3) accordingly. Given $\mathbf{I}$ and $g_1^1, \ldots, g_n^1$, constructing an optimal $\mathbf{G}^*$ is to maximize the following log posterior probability:

$$\begin{aligned}\mathbf{G}^* &= \arg\max_{\mathbf{G}} \log P(\mathbf{g}_1, \ldots, \mathbf{g}_N|g_1^1, \ldots, g_n^1, \mathbf{I}) \\ &= \arg\max_{\mathbf{G}} \sum_{i=1} \sum_{t=t_i^s} \log P(g_i^{t+1}|g_i^t, I^t, I^{t+1}) \\ &\quad + \sum_{i=n+1} \log p(g_i^{t_i^s}) \; s.t. \;\; \forall \mathbf{o} \;\; g_i^{t+1} = \Psi(g_i^t, \mathbf{o}).\end{aligned} \tag{8}$$

After initializing the grid partition in the first frame, we can perform exhaustive search to solve this problem. However, exhaustive search is computationally expensive due to enormous candidates. Specifically speaking, to find out an optimal solution, we have to determine the grid displacements of each frame for each grid flow. For example, let $[-u_x, u_x]$ and $[-u_y, u_y]$ denote the ranges of horizontal and vertical displacements, respectively. The total number of candidates is around $O(N \cdot (2u_x)^{N_T} \cdot (2u_y)^{N_T})$ given a video with N_T frames containing N grids flows, which is computationally prohibitive to find the optimal solution.

B. Simplified Grid Flow Construction

We propose to simplify the grid flow construction by formulating it as finding the longest path in a directed graph. The simplification consists of three operations as follows:

(a) From I^t to its following frame $I^{t'}$, we estimate a *dominant* displacement for all grids in I^t, rather than the individual displacements of each grid. With the estimated dominant displacement and g_i^t, we locate $g_i^{t'}$ in $I^{t'}$, and align g_i^t and $g_i^{t'}$ to construct grid flows. We then refine the grid alignments for those grids whose displacement is inconsistent with the dominant displacement.

(b) Given I^t, we employ the motion vectors of grids to estimate the grids for $I^{t+m} (m > 1)$. We hence generate the grids only for a subset of frames in a tracking manner. It is unnecessary to generate grids frame by frame, since camera/object motion is typically continuous, making the relocation of content slow and continuous between consecutive frames.

(c) We further restrict the range of dominant displacements, such that we can select key-frames in a reasonable motion range based on grid tracking. Specifically, from the current key-frame I^t to the next one I^{t+m}, the horizontal dominant displacement of grids from I^t is $m_1 \cdot w$, and the vertical

[1]For the sake of simplicity, we apply the translation motion model. More complex motion models can be utilized, but at the risk of increased complexity.

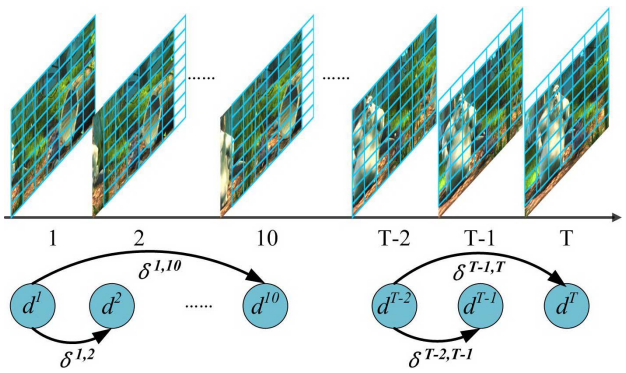


Fig. 4. Illustration of building a directed graph for the simplified model of grid flow construction.

displacement is $m_2 \cdot h$, where m_1 and m_2 are non-zero integers, and w and h respectively denote the width and height of the grid (See Fig. 4). Through such restriction, compared to I^t, the next selected I^{t+m} would contain new content which can be covered by at least a colume/row of grids, and the distance between two key-frames m is usually much greater than one. Hence, the number of key-frames is much smaller than the total number of video frames, whereas key-frames can well summarize the major video contents.

Through the above simplifications, grid flow construction is reduced to be the problem of simultaneously selecting key-frames and constructing sub-grid-flows via tracking over key-frames. Let $\mathbf{Q} = \{\mathbf{q}_k\}$ denote the set of all possible values of dominate displacement, where $\mathbf{q}_k$ is the k-th value, $\mathbf{R} = \{r_n\}$ the set of key-frames, where r_n indexes the n-th key-frame, $\bar{\mathbf{o}}_i^{r_n}$ the dominate displacement of $g_i^{r_n}$ from I^{r_n} to $I^{r_{n+1}}$, $\bar{\mathbf{G}}$ the sub-grid-flow set in key-frames. Equation (8) can be rewritten as

$$\begin{aligned}\{\mathbf{R}, \bar{\mathbf{G}}\}^* &\approx \arg\max_{\mathbf{G}} \sum_i \sum_{r\in\mathbf{R}} \log P(g_i^{r_{n+1}} | g_i^{r_n}, I^{r_n}, I^{r_{n+1}}) \\ &= \arg\max_{\mathbf{G}} \sum_{r\in\mathbf{R}} \sum_{i\in\mathbf{u}^r} \log P(g_i^{r_{n+1}} | g_i^{r_n}, I^{r_n}, I^{r_{n+1}}) \\ &\begin{cases} g_i^{r_{n+1}} = \Psi(g_i^{r_n}, \bar{\mathbf{o}}_i^{r_n}); \quad \bar{\mathbf{o}}_i^{r_n} \in \mathbf{Q} \\ \bar{\mathbf{o}}_i^{r_n} = \bar{\mathbf{o}}_j^{r_n}; \quad i, j \in \mathbf{u}^{r_n} \\ r_{n+1} - r_n > 2, \end{cases} \end{aligned} \tag{9}$$

where $\mathbf{u}^{r_n}$ denotes the set of the indices of all grids in I^{r_n}.

Building directed graph: As shown in Fig. 4, we construct a directed acyclic graph $\mathbf{\Omega}$, where each node represents a frame, and arc $a^{k,t,t'}$ indicates a grid moving by the dominant displacement $\mathbf{q}_k$ from I^t to $I^{t'}$. Let $\delta_k^{t,t'}$ denote the weight of $a^{k,t,t'}$, which represents the probability that all grids move by the dominant displacement $\bar{\mathbf{o}}_i^t = \mathbf{q}_k$ from I^t to $I^{t'}$:

$$\begin{aligned}\delta_k^{t,t'} &= \sum_{i\in\mathbf{u}^t} \log P(g_i^{t'} | g_i^t, I^t, I^{t'}) \\ s.t. \quad & g_i^{t'} = \Psi(g_i^t, \bar{\mathbf{o}}_i^t); \quad \bar{\mathbf{o}}_i^t = \mathbf{q}_k. \end{aligned} \tag{10}$$

We reformulate (9) in terms of the arcs' weights as

$$\begin{aligned}\{\mathbf{R}, \bar{\mathbf{G}}\}^* &\approx \arg\max_{\mathbf{R},\bar{\mathbf{G}}} \sum_{r\in\mathbf{R}} \sum_{i\in\mathbf{u}^r} \log P(g_i^{r_{n+1}} | g_i^{r_n}, I^{r_n}, I^{r_{n+1}}) \\ &= \arg\max_{\mathbf{R},\bar{\mathbf{G}}} \sum_{r\in\mathbf{R}} \delta_k^{r_n, r_{n+1}}. \end{aligned} \tag{11}$$

Algorithm 1 Grid Flow Construction Algorithm

Input:
Graph $\mathbf{\Omega}$.
Output:
Index of key-frames: $\mathbf{R}$,
Grid flow set on key-frames: $\bar{\mathbf{G}}$

```
initialize R = ∅
t = 1
L = 0
while t ≤ T do
    select neighboring frames of I^t: I^n = {I^{t'}}_{t'=t}^{t+m}
    find I^{t'} ∈ I^n, q_k ∈ Q that maximize L = L + δ_k^{t,t'}
    R = R ∪ t'
    determine grids on I^{t'}: g_i^{t'} = Ψ(g_i^t, ō_i^t),   ō_i^t = q_k
    find g^{t'*} that maximize P(g^{t'} | I^t, I^{t'}, g_i^t)
    g_i = g_i ⋃ g^{t'*}
    L = L + δ_k^{t,t'}
    t = t'
end while
{R, Ḡ}*
```

As a result, the grid flow construction is formulated as a problem of finding the longest path in the directed graph.

Removing unnecessary arcs: To improve searching efficiency, we remove unnecessary arcs. We only connect arcs from I^t to a few following frames $I^{t'}$, where $t+a < t' \leq t+b$, and a and b are set to be 2 and 22 empirically. In addition, between two nodes, we keep the arc with the highest weight and remove the rest, since the arc with the highest weight is most likely the optimal one.

Refining grid alignments: The displacements of some grids may be inconsistent with the dominant displacement $\bar{\mathbf{o}}_i^t$. Consequently, if we directly align g_i^t with $g_i^{t'}$ (i.e. $\mathbf{g}_i = \mathbf{g}_i \bigcup g_i^{t'}$) after we determine the grids on $I^{t'}$ according to $\bar{\mathbf{o}}_i^t$, inaccurate temporal alignments may be incurred. Therefore, these grid alignments need to be refined. Specifically, for g_i^t, we look for a grid on $I^{t'}$ which contains the most consistent content with the content of g_i^t by

$$g^{t'*} = \arg\max_{g^{t'}\in\xi(g_i^t)} P(g^{t'} | I^t, I^{t'}, g_i^t), \tag{12}$$

where $\xi(g_i^{t'})$ denotes the neighborhood of of $g_i^{t'}$.

Then we add $g^{t'*}$ to grid flow $\mathbf{g}_i$ and denote $g^{t'*}$ as $g_i^{t'}$.

The proposed graph is a directed graph with no negative-valued edges, which can be efficiently solved by the dynamic programming algorithm listed in Algorithm 1. Then, given the selected key-frames and estimated displacements, we can construct sub-grid-flows for key-frames. For the rest of non-key-frames, their sub-grid-flows can be constructed by grid interpolation as explained in Sec. V-B.

V. Grid-Flow-Guided Video Resizing

In this section, we describe how to use grid flows to resize individual frames. As detailed in Appendix A, we first resize

key-frames via quadratic programming. Then each non-key-frame is resized via grid interpolation based on the contents of its two nearest resized key-frames.

A. Optimization for Resizing Key-Frames

We formulate key-frame retargeting as a quadratic programming problem, for which we design an objective function to preserve the shape of salient objects. Moreover, in order to coherently deform an object between frames, we impose temporal coherence constraints and measure grid importance by considering temporal information.

The objective function: We employ the objective function proposed in our previous work [10] to optimally allocate resizing budgets of key-frames to the grids, according to grid-wise importance. To preserve the shape of salient objects, the objective function aims to maintain less deformation for the grids with higher importance as follows:

$$\min \sum_{i=1}^{N} \sum_{t \in \mathbf{R}} D(g_i^t) \cdot s_i^t, \tag{13}$$

where $D(\cdot)$ denotes the grid distortion metric, s_i^t the importance of grid g_i^t.

Grid Importance: The content importance of a single frame may not be meaningful for video. To measure the grid importance over a video clip is preferred. In our previous work [10], the importance of a grid is calculated from the current frame, and thus is limited to spatial importance. In this work, we propose to compute the importance of a grid by averaging the spatial importance values of those grids belonging to the same grid flow in key-frames as

$$s_i = \frac{1}{N_i} \sum_{g_i^t \in \mathbf{g}_i, t \in \mathbf{R}} \sum_{k,e \in g_i^t} im_{k,e}^t, \tag{14}$$

where N_i is the number of grids belonging to the same grid flow in key-frames, $im_{k,e}^t$ the spatial importance of the (k, e)-th pixel of I^t. In this work, we combine the saliency map obtained using the method in [36] and face detection results using the method in [37], to generate the importance map like [28], [29].

Distortion energy of a grid: We present a grid distortion energy function, which is different from that of conventional grid-based methods (e.g. [28] and [29]). Specifically, we restrict the shape of retargeted grids to be rectangular instead of arbitrary quadrilateral. We hence simply measure the distortion energy of a grid by the change of grid's aspect ratio due to retargeting. Furthermore, the rectangular restriction enables all retargeted grids in each row/column to have the same height/width. Thus, we represent the change of grid's aspect ratio in terms of the height of grid row and the width of grid column, rather than vertices. The distortion energy of g_i^t is defined as follows:

$$D(g_i^t) = \| w \cdot \widetilde{h}(l_i^t) - h \cdot \widetilde{w}(c_i^t) \|^2, \tag{15}$$

where w and h are the width and height of the grid, respectively, c_i^t and l_i^t the column and row indices of g_i^t, $\widetilde{w}(c_i^t)$ the resized width of grid column c_i^t, and $\widetilde{h}(l_i^t)$ the resized height of grid rows l_i^t.

Note, the energy function brings two advantages: (a) reducing the extent of freedom of grid deformation, avoiding severe shape distortions due to serious grid deformation [9], [11]; and (b) reducing the computational complexity of the optimization, since it leads to much fewer variables than those of vertex-based schemes (e.g. [28] and [29]).

Spatial coherence constraints: The size of each retargeted frame is constrained to meet the target size budget. Suppose a key-frame consists of $N_c \times N_r$ grid and the target size is $W' \times H'$, the following spatial coherence constraints are imposed:

$$\begin{cases} \sum_{c_i^t=1}^{N_c} \widetilde{w}(c_i^t) = W', & \forall t \in \mathbf{R} \\ \sum_{l_i^t=1}^{N_r} \widetilde{h}(l_i^t) = H', & \forall t \in \mathbf{R} \\ \widetilde{w}(c_i^t) > 0, & \forall c_i^t = 1, \ldots, N_c, \forall t \in \mathbf{R} \\ \widetilde{h}(l_i^t) > 0, & \forall l_i^t = 1, \ldots, N_r, \forall t \in \mathbf{R}. \end{cases} \tag{16}$$

Temporal coherence constraints: We further impose two temporal constraints to ensure temporal coherence. The first constrains the grids in each grid flow to be deformed coherently between key-frames. Given $\mathbf{g}_i$, the deformation of its grids satisfies

$$\begin{cases} \widetilde{w}(c_i^t) = \widetilde{w}(c_i^{t'}), & \forall t, t' \in \mathbf{R}, \forall g_i^t, g_i^{t'} \in \mathbf{g}_i \\ \widetilde{h}(l_i^t) = \widetilde{h}(r_i^{t'}), & \forall t, t' \in \mathbf{R}, \forall g_i^t, g_i^{t'} \in \mathbf{g}_i. \end{cases} \tag{17}$$

Due to the coarse granularity of grid flows, some grid flows may exhibit content inconsistency. Specifically, let $E_i^{t'}$ denote the evolution of content in grid g_i^t in frame $I^{t'}$. The temporally aligned grid $g_i^{t'}$ may cover only part of $E_i^{t'}$, and the rest of $E_i^{t'}$ may be covered by spatially adjacent grids of $g_i^{t'}$. To further ensure temporal coherence, we set the second temporal constraint to penalize the deformation discrepancy between g_i^t and the spatially adjacent grids of $g_i^{t'}$, according to the extent of content consistency between g_i^t and $g_i^{t'}$:

$$\begin{cases} b = P(g_i^{t'} | I^t, I^{t'}, g_i^t) \\ \frac{1}{\phi_1(b)} \le \frac{\widetilde{w}(c_i^t)}{\widetilde{w}(c_i^{t'} \pm 1)} \le \phi_1(b) & \forall t, t' \in \mathbf{R} \\ \frac{1}{\phi_2(b)} \le \frac{\widetilde{h}(l_i^t)}{\widetilde{h}(l_i^{t'} \pm 1)} \le \phi_2(b) & \forall t, t' \in \mathbf{R}, \end{cases} \tag{18}$$

where $P(g_i^{t'} | I^t, I^{t'}, g_i^t)$ measures the extent of content consistency between g_i^t and $g_i^{t'}$, $\phi_1(b) \ge 1$ and $\phi_2(b) \ge 1$ when $0 \le b \le 1$. We simply determine the spatially adjacent grids of $g_i^{t'}$ according to the row and column indices: given $g_i^{t'}$ at row $l_i^{t'}$ and column $c_i^{t'}$, its spatially adjacent grids are located at row $l_i^{t'} \pm 1$ or column $c_i^{t'} \pm 1$.

Global Optimal Solution: To determine the optimal retargeted grids, we minimize the objective function in (8) subject to the spatial coherence constraints in (16) and the temporal coherence constraints in (17) and (18). We employ the active-set method [38] to solve this optimization problem. With the initial guess $\frac{W'}{N_c}, \ldots, \frac{H'}{N_r}, \ldots$ satisfying the equality constraint, the nonlinear program can be solved iteratively to figure out the global optimal solutions in a feasible region. It is worth noting that the optimization is kind of convex quadratic programming. Specifically, the objective function is a quadratic one. The solutions satisfying the equality constraints and

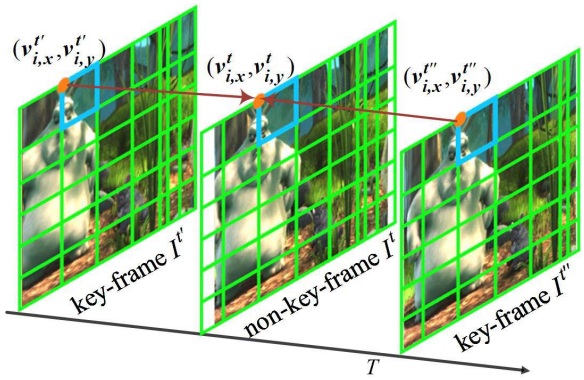


Fig. 5. Illustration of resizing non-key-frames, where all blue grids belong to grid flow $\mathbf{g}_i$.

inequality constraints form a convex set, since all the equality and inequality constraints are linear and can be regarded as concave functions. Therefore, the global optimal solution is obtained when a local solution is resolved.

Complexity Analysis: Since the Hessian matrix of the objective function is positive semidefinite, and the optimization is a convex programming problem, similar to linear programming, the complexity of our optimization depends on the number of the model parameters and the number of constraints. In other words, our optimization can be solved in polynomial time [39].

B. Grid Interpolation for Resizing Non-Key-Frames

Suppose that the content of a non-key-frame can be predicted from its nearest key-frames, we propose to retarget non-key-frames based on the retargeting results of the neighbor key-frames. Considering short time interval between two nearest key-frames, we assume the trajectory of grids moves linearly frame by frame between two key-frames. Thus, we employ linear interpolation to resize each grid in non-key-frames in between two key-frames.

Since some grids may not fully cover the non-key-frames, we propose to resize grids based on the location of their top-left vertices rather than widths/heights. As shown in Fig. 5, let $\tilde{v}_{i,x}^{t}$ and $\tilde{v}_{i,y}^{t}$ denote the x and the y indices of the top-left vertex of a resized grid g_i^t, respectively. Suppose I^t is a non-key-frame, $I^{t'}$ and $I^{t''}$ $(t' \leq t \leq t'')$ are its nearest key-frames. According to $g_i^{t'}$ and $g_i^{t''}$ we calculate the vertices of the retargeted version of g_i^t:

$$\begin{cases} \tilde{v}_{i,x}^{t} = \tilde{v}_{i,x}^{t'} + \frac{\tilde{v}_{i,x}^{t''} - \tilde{v}_{i,x}^{t'}}{t''-t'}(t-t') \\ \tilde{v}_{i,y}^{t} = \tilde{v}_{i,y}^{t'} + \frac{\tilde{v}_{i,y}^{t''} - \tilde{v}_{i,y}^{t'}}{t'-t''}(t-t'), \end{cases} \tag{19}$$

where $\tilde{v}_{i,x}^{t'} = \sum_{j=1}^{c_i^{t'}-1} \tilde{w}(j)$, $\tilde{v}_{i,y}^{t'} = \sum_{k=1}^{r_i^{t'}-1} \tilde{h}(k)$, and $\tilde{v}_{i,x}^{t''}, \tilde{v}_{i,y}^{t''}$ are calculated likewise.

Compared to those methods working out optimization to resize all frames, grid interpolation brings about two advantages: (a) the interpolation tends to smooth grid motion, securing continuous motion of pixels across retargeted frames to avoid jittering artifacts; (b) the interpolation drastically reduces the computational cost, since the grid interpolation is of low computational complexity and non-key-frames occupy a major proportion of a video.

VI. Experiment

We first validate that constructing grid flows and selecting key-frames can reduce the inconsistent content which have been constrained to coherently deform between frames; we then compare our method to the state-of-the-art approaches. The computational complexity will be evaluated as well.

A. Validations

The more regions/pixels involving different contents are constrained to be deformed coherently cross frames, the more shape preservation would be degraded. We show that the proposed key-frame-based grid flow can effectively address such problem.

To quantitatively assess temporal inconsistent content, an ideal approach is to extract video objects, and then to measure the area of regions/pixels that belong to different objects but are constrained to be coherently deformed between frames. As accurate object segmentation is still very challenging, we propose a practical approach to the quantitative assessment via two simplifications: (a) We divide the visual content of frames into two groups: salient content and non-salient content. This is motivated by the concern about whether contents of different visual importance are properly deformed in terms of shape preservation. (b) We determine the content saliency at the grid level. That is, a grid is considered as a salient one, if more than 25% of its area is salient; otherwise, as a non-salient grid. We propose to assess the temporal inconsistent content by counting in a frame the number of salient grids which are constrained to be deformed coherently with at least a non-salient grid in other frames and that of non-salient grids constrained to be deformed coherently with at least a salient grid in other frames. The count of inconsistent content is then normalized by dividing the total number of grids in the frame.

To study the temporal inconsistency, we select challenging videos containing significant motions. We manually annotate salient and non-salient regions for each frame as the ground truth.

Grid flow: To evaluate the effectiveness of our grid-flow-based method in mitigating inconsistent content, we assess the amount of inconsistent content generated by enforcing all grids in each grid flow to be deformed coherently. We name such temporal strategy as "global temporal strategy on grid flows" (GTS + GF). We compare GTS + GF with Motion-based Video Retargeting (MVR) [28] and Scalable and Coherent Video Resizing (SCVR) [29], which are the state-of-the-art global methods. Both MVR and SCVR divide each frame into grids independently, employ optical flow to align grids between frames, and then constrain each grid to be deformed coherently with the aligned grids in other frames in the similar way to GTS + GF.

As illustrated in Figs. 6 and 7, GTS + GF leads to less inconsistent content than MVR and SCVR. The main reason is that GTS + GF takes into account the contents of grids in the previous frame during grid flow construction. In contrast, the grid partitions of MVR and SCVR do not

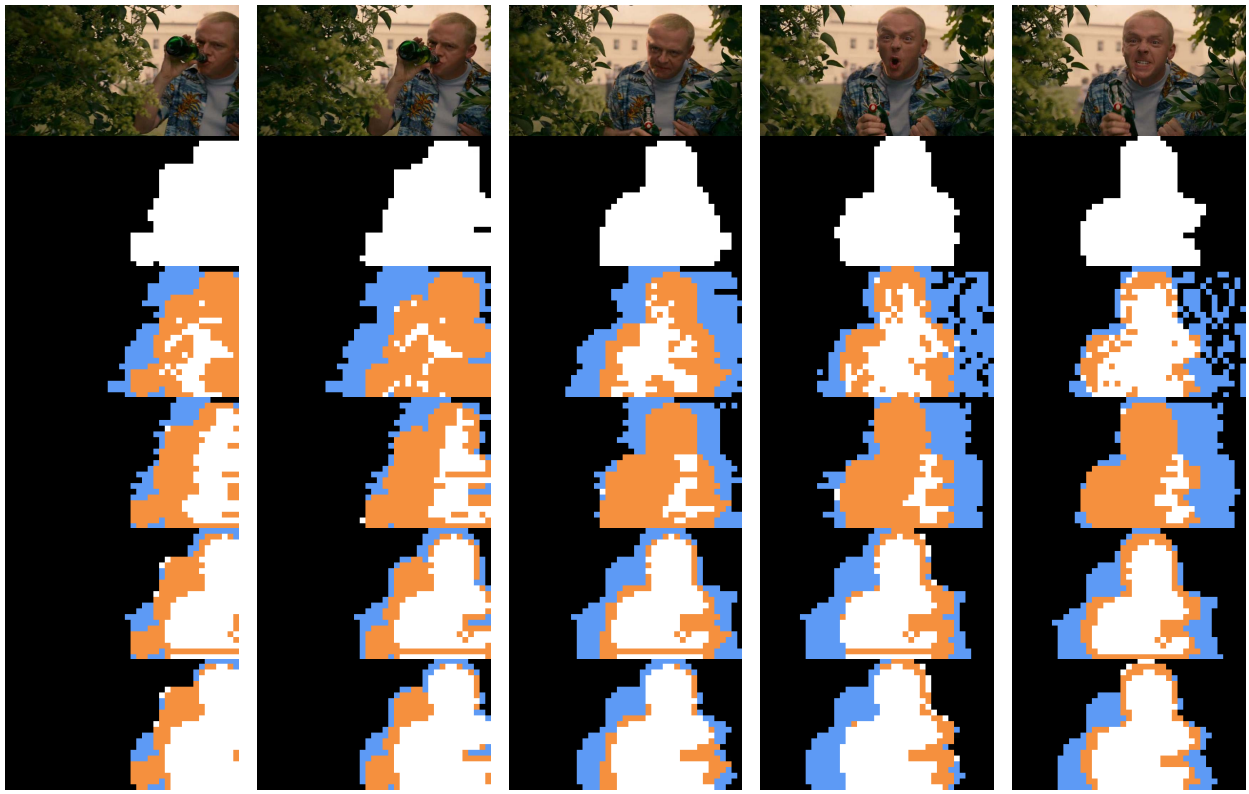


Fig. 6. Regions which are constrained to be deformed coherently but incurring at least one region containing inconsistent contents in other frames. A non-salient grid is colored blue, should it be constrained to be deformed coherently with at least a salient grid in other frames, and black otherwise. Likewise, a salient grid is colored orange, should it be constrained to be deformed coherently with at least a non-salient grid in other frames, and is colored white otherwise. Rows from top to bottom: (1) the original frame; (2) the ground-truth; (3) MVR [27]. (4) SCVR [29]. (5) GTS + GF. (6) KTS+GF (our method).

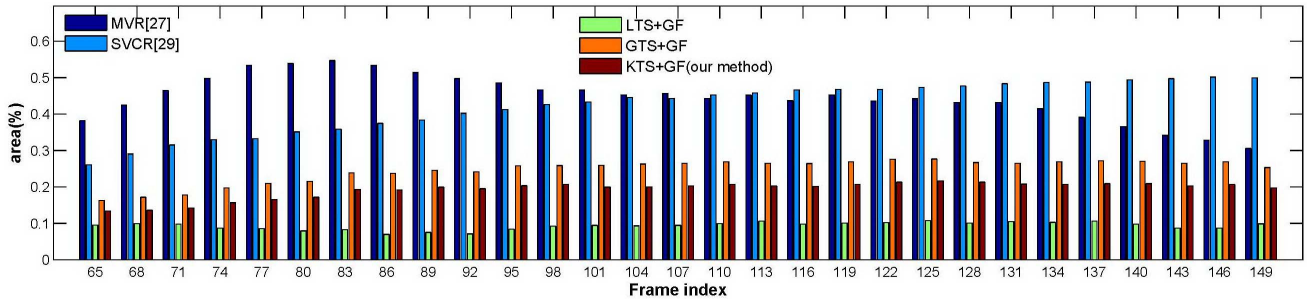


Fig. 7. Area of inconsistent contents which are constrained to deform coherently between frames.

consider the correlations between frames, leading to many incorrect alignments. That is, given a grid, its aligned grid contains inconsistent content partially.

Key-frames: To evaluate the effectiveness of our key-frame-based scheme in mitigating inconsistent content, we assess the amount of inconsistent content caused by enforcing the grids in each grid flow to be deformed coherently among key-frames. We name such strategy as "key-frames on grid flows" (KTS+GF). We compare KTS+GF with two other temporal strategies. One is GTS + GF. The other enforces that grids in each grid flow are deformed coherently among 25 consecutive frames, namely "local temporal strategy on grid flows" (LTS+GF).

As shown in Fig. 7, the inconsistent areas of KTS + GF and LTS + GF are much smaller than that of GTS + GF. KTS + GF and LTS + GF use fewer frames than GTS + GF, thereby reducing the possibility of grids with inconsistent contents being constrained to be deformed coherently. However, unlike KTS + GF, LTS + GF performs poorly in maintaining temporal coherence for videos with large motions.

B. Retargeting Results

For performance evaluation, we collect various videos with four considerations: (1) The test videos shall have enough content diversity; (2) The video frames shall contain salient objects/regions of reasonable sizes. To evaluate the effects of temporal retargeting strategies on shape preservation, frames are supposed to involve considerable non-salient regions. (3) Each test video shall cover different types of motions, such as camera motion, object motions, and simultaneous camera and object motions; (4) To challenge retargeting schemes in dealing with large motion videos, we select test videos with high motion activities for performance evaluation.

(a) (b) (c) (d) (e)

Fig. 8. Retargeting results for key-frames and non-key-frames: (a) the original frames, where key-frames are in the 1st column and non-key-frames are in the 2nd column. (b), (d) retargeted key-frames. (c), (e) retargeted non-key-frames.

In our experiments, each video is retargeted to 50% width of the source video. The complete experimental video results (including all comparison results) can be downloaded from our project website [40]. As demonstrated in [40] and Fig. 8, by retargeting key-frames and non-key-frames separately, our method can well preserve the shape of salient objects for both key-frames and non-key-frames, while coherently resizing objects across adjacent frames. (Readers are referred to [40] for subjective comparisons.)

The quality of video retargeting is subjective and dependent on user preference.[2] Like [18], [22], [28], we not only qualitatively compare our method with state-of-the-art methods, but also conduct user study which ranks the compared methods via paired comparisons [41].

For comprehensive comparisons, we have selected five representative methods including Streaming Video Retargeting (SVR) [18], Mosaic-Guided Scaling (MGS) [26], Motion-aware video Retargeting (MAR) [27], Motion-based video Retargeting (MVR) [28], and Scalable and Coherent Video Resizing (SCVR) [29]. Note that, MVR and SCVR are hybrid methods, which combine grid-based warping with cropping. So we first compare our method with SVR, MGS and MAR, and then with MVR and SCVR.

Qualitative comparisons: We compare the performances of these methods in terms of temporal coherence and spatial shape preservation of salient objects, respectively.[3]

[2]There is no widely-accepted objective criterion to quantitatively evaluate retargeting results to the best of our knowledge.

[3]Due to space limit, we are unable to list numerous frames for illustrating the comparison of retargeting effects. Readers are referred to the complete set of demo video clips [40] for better feelings on the qualitative comparisons and the subsequent subjective description.

Compared with SVR, MGS and MAR, our method achieves the best performance in maintaining temporal coherence. In contrast, SVR can maintain temporal coherence for videos with small motions but is often ineffective for videos with large motions. For example, as shown in Fig. 9 and the demonstration video clips in [40], SVR shrinks the region of left background and then stretches it to the normal shape for the video *Restaurant*. This is because SVR uses local temporal information to ensure temporal coherence, and such temporal information fails to cover the duration of the camera motion. In contrast, although MGS and MAR can achieve satisfactory temporal coherence for most test videos using global temporal information, they produce waving artifacts on the videos with homogeneous background (e.g. video *Call* in Fig. 9 and [40]), since their employed feature-based motion estimation often yields inaccurate temporal alignments on such videos. Our method maintains temporal coherence well over all these videos, since the key-frames can summarize the temporal information of a video and accordingly, the corresponding content between frames can be properly aligned by the grid flows.

Note, MAR, MGS and SVR all distort salient objects, though the frames contain enough non-salient regions. For example, as shown in Fig. 9, the man is seriously distorted by MAR, SVR and MGS for the video *Restaurant*, in spite of large area of non-salient background. This is because these methods improperly constrain quite a few regions/pixels containing inconsistent content to be deformed coherently between frames. For example, SVR attempts to align temporally adjacent pixels and constrain them to be coherently deformed, while these adjacent pixels involve distinct video content in videos with large motions. Our method achieves the best performance in shape preservation of salient objects,

(a) (b) (c) (d) (e)

Fig. 9. Visual quality comparison: (a) the original frame, and the retargeted frames using (b) SVR [18], (c) MAR [28], (d) MGS [26], (e) Our results. Rows from top to bottom, the original frames are from the videos which we name: *Call*, *Tree*, *Restaurant*. (Refer to our demonstration video clips [40]).

since our key-frame-based grid flows avoid deforming such inconsistent content. (See the comparison results in [40].)

Compared with MVR and SCVR, both our method and SCVR achieve good temporal coherence for all these videos, but MVR fails in some video frames. For example, as shown in Fig. 10 and [40], MVR distorts the motorcycle man's body between frames for the video *Motorcycle*, and causes flickering artifacts in the chest and right colorful box for the video *Toy*. The reason is that the temporal strategy of MVR is sensitive to the errors of motion estimation.

Thanks to cropping, both MVR and SCVR preserve the shapes of salient objects quite well. However, the cropping may discard some context or part of salient objects. For example, as shown in Fig. 10, MVR crops out the right electric pole for the video *Motorcycle*; SCVR crops out the right chair for the video *Airship* as well as the rightmost colored box for the video *toy*. Moreover, cropping often introduces false camera motion. For example, there contains no camera motion in the video *Toy* (see Fig. 10 and [40]), but the retargeted videos by SCVR show a fake camera motion tracking towards the right. Without cropping, our method achieves comparable or even better shape preservation than MVR and SCVR, since our method yields fewer regions that contain inconsistent content but are enforced to deformed coherently.

Subjective user study: We conduct a user study to subjectively evaluate the performance. We invite 50 subjects of different ages and professions to participate in the user study. For each subject, we show a batch of original videos, as well as the pairs of retargeted videos generated by different video methods for clear comparison. To avoid personal bias, the retargeted videos are presented in a random order, so that these subjects have no idea of which method produces the displayed retargeted videos. Each subject is required to prefer one of them. By means of user preference and paired comparisons, the performance is measured quantitatively.

TABLE I
PREFERENCES OF 50 SUBJECTS FOR SVR, MAR, MGS AND OUR METHOD

	Ours	SVR	MAR	MGS	Total
Our method	-	182	173	177	532
SVR	68	-	123	118	309
MAR	77	127	-	95	299
MGS	73	132	155	-	360

TABLE II
PREFERENCES OF 50 SUBJECTS FOR MVR, SCVR AND OUR METHOD

	Ours	MVR	SCVR	Total
Our method	-	230	245	475
MVR	120	-	165	285
SCVR	105	185	-	290

We compare our method with SVR, MAR and MGS over five different videos and perform 6 paired comparisons for each video. In total, we receive $5 \times 6 \times 50 = 1500$ answers and each method is pairwise compared by $3 \times 5 \times 50 = 750$ times. Table I shows the count of preferred retargeted videos of each method by paired comparisons. Note that in Table I and II each entry means that the method in row r is ranked better n times than the method in column c. Table I shows that our method outperforms the other three methods in 70.9% (532/750) of the paired comparisons. The breakdown is that our method is preferred in 72.8% (182/250) with SVR, in 69.2% (173/250) comparisons with MAR, and in 70.8% (177/250) with MGS.

We also compare our method with MVR and SCVR over seven different videos. Hence, we generate 3 paired comparisons for each video. In total, we receive $3 \times 7 \times 50 = 1050$

(a) (b) (c) (d) (e)

Fig. 10. Visual quality comparison: (a) the original frames, and the retargeted frames using (b) SVR [18]. (c) MVR [28]. (d)SCVR [29]. (e) Our results. Rows from top to bottom, the original frames are from the videos which we name: *Airship*, *Motorcycle*, *Toy*. (Refer to our demonstration video clips [40]).

answers and each methods is pairwise compared by $2 \times 7 \times 50 = 700$ times. Table II shows the results of user study, where the results shows clear preference for our method. Our method is preferred in 67.9% (475/700) paired comparisons with other two methods, including preference of 65.7% (230/350) with MVR and 70.0% (245/770) with SCVR.

C. Memory and Runtime Cost

We have implemented our retargeting system on a laptop personal computer with Intel Duo 2.26 GHz CPU, 2GB memory. To show the efficiency, we use a single core without any parallel processing. We do not employ any hardware (e.g., GPU) based acceleration. We compare our method with MVR [28] and SCVR [29]. As reported in [28] and [29], MVR and SCVR were implemented on better hardware configurations: MVR on a PC with Duo 2.33 GHz CPU and Nvidia GTX 285 graphics card; SCVR on a PC with 4 Core 2.66 GHz CPU and 8 GB memory, using CPU-based parallel processing.

The memory and runtime costs are mainly consumed by the grid flow construction and the optimization for key-frame resizing. As shown in Table III, the grid flow construction consumes low memory and runtime cost for a video of 221 frames. Our grid flow construction is not memory-demanding since there is no need to load the entire video at once. Instead, we store the color histograms of grids from a small number of frames. Thus, the memory cost depends on the number of grids of the first frame, the number of frames, and the dimension of histogram. In addition, the time complexity of grid flow construction is $O(T \times N_c \times N_r)$, where T is the number of frames, $N_c \times N_r$ the number of grids of the first frame.

In key-frame resizing, both runtime cost and memory cost are spatial resolution independent, whereas they are proportional to the number of variables. We list the memory and runtime cost in Table IV. Our optimization consumes

TABLE III

MEMORY AND RUNTIME COST CONSUMED BY GRID FLOW CONSTRUCTION. THE BOUNDED NUMBER IS 221, THE DIMENSION OF COLOR HISTOGRAM IS 70

Grid division	14×34	29×69	57×137
Time	3s	6s	19s
Memory	29MB	40MB	61MB

TABLE IV

MEMORY AND RUNTIME COST CONSUMED BY OPTIMIZATION

Grid division	time			memory		
	MVR	SCVR	our	MVR	SCVR	our
14×34	67s	2s	0.012s	225MB	22MB	5.8MB
29×69	219s	10s	0.028s	880MB	100MB	7.6MB
57×137	575s	41s	0.320s	3.5GB	432MB	14.1MB

lower memory and runtime cost than MVR and SCVR. This is attributed to the significantly reduced number of variables for optimization. Rather than performing optimization on all frames, we optimize the grids in key-frames, which involve only 8%-10% of video frames. Furthermore, our method reduces the number of the optimization variables to about $N_B \times (N_c + N_r)$, where N_B is the number of key-frames. In contrast, performing optimization on all vertices for all frames would significantly increase the number of variables to $N_T \times (N_c \times N_r)$, where N_T is the total number of frames.

For HD videos, to achieve high visual quality with reasonable computation and memory cost, we can first use a coarse grid partition for the grid flow construction, and then further partition the coarse grids into fine grids for the optimization process, where we resolve the displacement of fine grids based on the displacement of corresponding coarse grids. For example, suppose the coarse grid partition of the 1st frame is 20×20 (i.e. 20 rows and 20 columns) in grid flow construction, and 40×40 fine grid partition is applied in the

optimization. The displacement of each fine grid is half of the displacement of the corresponding coarse grid. Overall, our prototype achieves about 50 fps for HD video, whereas MVR and SCVR achieve about 6 fps and 75 fps, respectively, on a 688×288 resolution video. Since the time complexity of MVR and SCVR is resolution dependent, lower frame rate with MVR and SCVR would be expected for HD video. In addition, SCVR utilizes CPU-based parallel processing, but our implementation does not.

VII. Conclusion

We proposed a novel video retargeting method that divides a video into grid flows. Our method utilizes grid flows to select key-frames, and then resizes these key-frames via quadratic programming to minimize the risk that grids containing inconsistent content are coherently deformed across frames. The remaining non-key-frames are subsequently resized by low-complexity grid interpolation based on the contents of the nearest retargeted key-frames. With the proposed grid-flow-guided retargeting for key-frames and grid-interpolation for non-key-frames, our method simultaneously achieves good temporal coherence and shape preservation of salient objects at low computational cost. Our experiments show that the proposed method is effective and efficient to handle videos of various types and resolutions.

Appendix

Algorithm 2 Grid-Flow-Guided Video Resizing Algorithm

Input:
video $\mathbf{I} = \{I_i^t\}_{i=1}^T$
grid flow set $\mathbf{G}$
index of key-frames R.
Output:
retargeted video $\tilde{\mathbf{I}}$
initialize temporal constraints
initialize spatial constraints
for $t \in \mathbf{R}$ **do**
  determine location of grid flows in frame I^t
  set temporal constraints for I^t.
  set spatial constraints for I^t.
end for
find $\tilde{w}(l)$, $\tilde{h}(r)$ to minimize the objective function subject to the temporal spatial coherence constraints
for $t \notin \mathbf{R}$ **do**
  find two nearest key-frames I^{t_k}, I^{t_j}
  calculate $\tilde{v}_{i,x}^t$ by interpolation based on $\tilde{v}_{i,x}^{t_k}$, $\tilde{v}_{i,x}^{t_j}$.
  calculate $\tilde{v}_{i,y}^t$ by interpolation based on $\tilde{v}_{i,y}^{t_k}$, $\tilde{v}_{i,y}^{t_j}$.
end for
reconstruct $\tilde{\mathbf{I}}$ according to $\tilde{w}(l)$, $\tilde{h}(r)$, $\tilde{v}_{i,x}^t$, $\tilde{v}_{i,y}^t$.

References

[1] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Syst.*, vol. 9, no. 4, pp. 353–364, Oct. 2003.

[2] H. Liu, X. Xie, W.-Y. Ma, and H. Zhang, "Automatic browsing of large pictures on mobile devices," in *Proc. 11th ACM Int. Conf. Multimedia*, Nov. 2003, pp. 148–155.

[3] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proc. 16th Annu. ACM Symp. User Inter. Softw. Technol.*, May 2003, pp. 95–104.

[4] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Apr. 2006, pp. 771–780.

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

[6] A. Mansfield, P. Gehler, L. Van Gool, and C. Rother, "Scene carving: Scene consistent image retargeting," in *Proc. ECCV*, Sep. 2010, pp. 143–156.

[7] 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, p. 118, Dec. 2008.

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

[9] B. Li, L.-Y. Duan, J. Wang, J. Chen, R. Ji, and W. Gao, "Grid-based retargeting with transformation consistency smoothing," in *Proc. 17th Int. Multimedia Model. Conf.*, Jan. 2011, pp. 12–24.

[10] B. Li, Y. Chen, J. Wang, L.-Y. Duan, and W. Gao, "Fast retargeting with adaptive grid optimization," in *Proc. IEEE ICME*, Jul. 2011, pp. 1–4.

[11] D. Panozzo, O. Weber, and O. Sorkine, "Robust image retargeting via axis-aligned deformation," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 229–236, May 2012.

[12] P.-Y. Laffont, J. Y. Jun, C. Wolf, Y.-W. Tai, K. Idrissi, G. Drettakis, *et al.*, "Interactive content-aware zooming," in *Proc. Graph. Inter.*, 2010, pp. 79–87.

[13] C.-H. Chang, "A line-structure-preserving approach to image resizing," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 1075–1082.

[14] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Trans. Graph.*, vol. 28, no. 3, p. 23, Aug. 2009.

[15] W. Dong, N. Zhou, J.-C. Paul, and X. Zhang, "Optimized image resizing using seam carving and scaling," *ACM Trans. Graph.*, vol. 29, no. 5, pp. 1–10, Dec. 2009.

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

[17] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–6.

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

[19] P. Greisen, M. Lang, S. Heinzle, and A. Smolic, "Algorithm and VLSI architecture for real-time 1080p60 video retargeting," in *Proc. 4th ACM SIGGRAPH/Eurograph. Conf. High Perform. Graph.*, 2012, pp. 57–66.

[20] 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, Oct. 2008.

[21] Y. Niu, F. Liu, X. Li, and M. Gleicher, "Warp propagation for video resizing," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2010, pp. 537–544.

[22] B. Yan, K. Sun, and L. Liu, "Matching area based seam carving for video retargeting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 302–310, Feb. 2013.

[23] W.-L. Chao, H.-H. Su, S.-Y. Chien, W. Hsu, and J.-J. Ding, "Coarse-to-fine temporal optimization for video retargeting based on seam carving," in *Proc. IEEE ICME*, Jul. 2011, pp. 1–6.

[24] L. Shi, J. Wang, L. Duan, and H. Lu, "Consumer video retargeting: Context assisted spatial-temporal grid optimization," in *Proc. 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp. 301–310.

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

[26] 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.

[27] 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, p. 127, Dec. 2009.

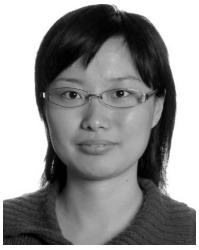
[28] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," *ACM Trans. Graph.*, vol. 29, no. 4, p. 90, Jul. 2010.

[29] 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, p. 88, Aug. 2011.

[30] Y. Hu and D. Rajan, "Hybrid shift map for video retargeting," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2010, pp. 577–584.

[31] J. Sun and H. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *Proc. IEEE ICCV*, Nov. 2011, pp. 1511–1518.

[32] T. Deselaers, P. Dreuw, and H. Ney, "Pan, zoom, scan—Time-coherent, trained automatic video cropping," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2008, pp. 1–8.
[33] G. Hua, C. Zhang, Z. Liu, Z. Zhang, and Y. Shan, "Efficient scale-space spatiotemporal saliency tracking for distortion-free video retargeting," in *Proc. 9th ACCV*, 2010, pp. 182–192.
[34] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, Oct. 2006, pp. 241–250.
[35] X. Fan, X. Xie, H. Q. Zhou, and W. Y. Ma, "Looking into video frames on small displays," in *Proc. 11th ACM Int. Conf. Multimedia*, Nov. 2003, pp. 247–250.
[36] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.
[37] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[38] J. Nocedal and S. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer-Verlag, 2006.
[39] M. Kozlov, S. Tarasov, and L. Khachiyan, "Polynomial solvability of convex quadratic programming," *USSR Comput. Math. Math. Phys.*, vol. 20, no. 5, pp. 223–228, 1980.
[40] (2013). *Video Retargeing Project* [Online]. Available: http://vipl.ict.ac.cn/members/bli
[41] H. A. David, *The Method of Paired Comparisons*. London, U.K.: Griffin, 1963.



Bing Li received the B.S. degree in computer science from Ji'nan University, Guangzhou, China, in 2009. She is currently pursuing the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. She is with the Institute of Digital Media, School of EE & CS, Peking University, Beijing. Her current research interests include image/video processing.

Ling-Yu Duan (M'06) received the Ph.D. degree in information technology from the University of Newcastle, Australia, in 2007, the M.Sc. degree in computer science from the National University of Singapore, Singapore, and the M.Sc. degree in automation from the University of Science and Technology of China, Hefei, China, in 2002 and 1999, respectively. Since 2008, he has been with Peking University, Beijing, China, where he is currently an Associate Professor with the School of Electrical Engineering and Computer Science. Dr. Duan is leading the group of visual search with the Institute of Digital Media, Peking University. Since 2012, Dr. Duan is the Deputy Director with the Rapid-Rich Object Search (ROSE) Laboratory, a joint laboratory between the Nanyang Technological University, Singapore, and the Peking University, China, with a vision to create the largest collection of structured domain object database in Asia, and to develop rapid and rich object mobile search. Before that, he was a Research Scientist with the Institute for Infocomm Research, Singapore, from 2003 to 2008. His current research interests include the areas of visual search and augmented reality, multimedia content analysis, and mobile media computing. He has authored more than 90 publications in these areas.

Jinqiao Wang received the B.E. degree from Hebei University of Technology, China, and the M.S. degree from Tianjin University, China, and the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, in 2001, 2004, and 2008, respectively. He is currently an Associate Professor with the Chinese Academy of Sciences. His current research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Rongrong Ji is currently a Professor, the Director of the Intelligent Multimedia Technology Laboratory, and the Dean Assistant of School of Information Science and Engineering with Xiamen University. His scholarly work mainly focus on innovative technologies for multimedia signal processing, computer vision, and pattern recognition, with over 100 paper published in international journals and conferences. He has received the ACM Multimedia Best Paper Award and the Best Thesis Award of Harbin Institute of Technology. He serves as an Associate/Guest Editor for international journals and magazines such as Neurocomputing, Signal Processing, Multimedia Tools and Applications, and the IEEE Multimedia Magazine and Multimedia Systems. He serves as a program committee member for several tier-1 international conferences. He is a member of ACM.

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, Institute of Communications Engineering, NTHU. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. Prior to joining academia, he worked for the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, from 1992 to 2000. His current research interests include image and video processing, and video networking. Dr. Lin is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE MULTIMEDIA, and the *Journal of Visual Communication and Image Representation*. He is an Area Editor of EURASIP Signal Processing: Image Communication. He serves as a member of the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA. He is currently the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society. He served as the Technical Program Co-Chair of the IEEE International Conference on Multimedia & Expo (ICME) in 2010, and the Special Session Co-Chair of the IEEE ICME in 2009. He is a Distinguished Lecturer of Asia-Pacific Signal and Information Processing Association. His paper won the Young Investigator Award presented by VCIP 2005. He received the Young Faculty Awards presented by CCU in 2005, and the Young Investigator Awards presented by National Science Council, Taiwan, in 2006.

Wen Gao (M'92–SM'05–F'08) received the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo, in 1992, and a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, in 1993. From 1994 to 1995, he was a Visiting Professor with the AI Laboratory, Massachusetts Institute of Technology. He is currently a Professor with the School of Electronic Engineering and Computer Science, Peking University, Peking, China, and a Professor in computer science with the Harbin Institute of Technology. He is the Honor Professor in computer science with the City University of Hong Kong, Hong Kong, and the External Fellow of International Computer Science Institute, University of California, Berkeley, CA, USA. His current research interests include signal processing, image and video communication, computer vision, and artificial intelligence.