Context-Aware Video Retargeting via Graph Model

Zhan Qu, Jinqiao Wang, *Member, IEEE*, Min Xu, *Member, IEEE*, and Hanqing Lu, *Senior Member, IEEE*

***Abstract*—Video retargeting is a crowded but challenging research area. In order to maximally comfort the viewers' watching experience, the most challenging issue is how to retain the spatial shape of important objects while ensure temporal smoothness and coherence. Existing retargeting techniques deal with these spatial-temporal requirements individually, which preserve the spatial geometry and temporal coherence for each region. However, the spatial-temporal property of the video content should be context-relevant, i.e., the regions belonging to the same object are supposed to undergo uniform spatial-temporal transformation. Regardless of the contextual information, the divide-and-rule strategy of existing techniques usually incurs various spatial-temporal artifacts. In order to achieve satisfactory spatial-temporal coherent video retargeting, in this paper, a novel context-aware solution is proposed via graph model. First, we employ a grid-based warping framework to preserve the spatial structure and temporal motion trend at the unit of grid cell. Second, we propose a graph-based motion layer partition algorithm to estimate motions of different regions, which simultaneously provides the evaluation of contextual relationship between grid cells while estimating the motions of regions. Third, complementing the salience-based spatial-temporal information preservation, two novel context constraints are encoded for encouraging the grid cells of the same object to undergo uniform spatial and temporal transformation, respectively. Finally, we formulate the objective function as a quadratic programming problem. Our method achieves a satisfactory spatial-temporal coherence while maximally avoiding the influence of artifacts. In addition, the grid-cell-wise motion estimation could be calculated every few frames, which obviously improves the speed. Experimental results and comparisons with state-of-the-art methods demonstrate the effectiveness and efficiency of our approach.**

***Index Terms*—Context-aware, grid graph model, spatial-temporal correlation, video retargeting.**

I. Introduction

VIDEO retargeting is to adapt videos to various display devices with clear and smooth imaging quality. Given a display device screen, the original video is adapted to a suitable version in terms of scale and aspect ratio. With the development of multimedia and Internet techniques, video retargeting becomes a crowded research area due to the proliferation of video data presented across various digital display platforms, from TVs, PCs, PDAs, to cell phones. The key issue of high-quality video retargeting is retaining the temporal smoothness and coherence as well as avoiding spatial shape distortion and maximally comfort the video viewers' watching experience.

Most existing retargeting techniques accomplish this purpose by preserving the spatial and temporal information of different regions, respectively, which are based on the spatial salience map and temporal motion estimation (e.g., flow estimation method). However, for video retargeting, this respective strategy usually incurs two inevitable problems. First, the salience map across same object is usually nonhomogeneous, which results in nonuniform spatial deformation in retargeting. This leads to noticeable damage to object geometry. Second, the regional motions of same object are generally continuous, and yet the existing approaches preserving the temporal coherence of regions individually cause the regions belonging to the same object undergo inconsistent temporal transformation and results in the corresponding temporal artifacts, such as background waving and foreground flicker. These spatial-temporal artifacts degenerate the quality of retargeting significantly.

Thus, an ideal spatial-temporal coherent retargeting framework should embody the awareness of contextual relationship, that is to say, regions belonging to the same object should experience similar spatial-temporal transformation. However, the object identification and segmentation is still a challenging issue in computer vision, and the direct application of these algorithms is neither practical nor economic for boosting the retargeting performance. Fortunately, it is easy to notice that the temporal motion and spatial content in video are correlated due to the following facts: 1) compared with the regions belonging to different objects, those belonging to same object are visually similar and 2) the regional motions across same object are usually continuous and have the similar direction and intensity. Correspondingly, the neighboring regions with similar visual content and motion more likely belong to the same object. Therefore, we believe that these correlations are helpful for modeling the contextual relationship between regions, which would improve the quality of retargeting significantly.

In this paper, we propose a novel graph model-based context-aware video retargeting framework. First, we employ a grid based warping algorithm to preserve the spatial structure and temporal motion of source videos at the unit of grid cell. Compared with others, this method provides a smoother imaging quality and a higher operation efficiency. Second, instead of the pixel-wise optical flow, we propose a graph-based motion layer partition algorithm to estimate the grid-cell-wise motion. Comparing with pixel-wise flow estimation, the coarse grid and the consideration on visual similarity of neighboring regions help to suppress the vulnerability to noise disturbance and produce

Manuscript received September 18, 2012; revised December 19, 2012, February 27, 2013; accepted February 28, 2013. Date of publication June 11, 2013; date of current version October 11, 2013. This work was supported in part by 973 Program under Grant 2010CB327905 and the National Natural Science Foundation of China under Grants 61070104, 61003161, 60905008, and 61273034. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Monica Aguilar.

Z. Qu, J. Wang, and H. Lu are with the National Laboratory of Pattern Recognition, Institute of Automations, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zqu@nlpr.ia.ac.cn; jqwang@nlpr.ia.ac.cn; luhq@nlpr.ia.ac.cn).

M. Xu is with iNEXT, School of Computing and Communications, University of Technology, Sydney 2007, Australia (e-mail: min.xu25@gmail.com).

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/TMM.2013.2267727

robust and reliable motion estimation. More importantly, due to the structure of a grid graph model, the motion information and visual similarity of regions are reflected by the weights of different edges respectively, which just corresponds to the definition of spatial-temporal correlation. Consequently, in addition to motion estimation, the graph model provides the measurement of contextual relationship between regions simultaneously, which factually reflects the possibility that the neighboring regions belong to same object. This means the graph method can be fused into our context aware framework seamlessly. Thirdly, complementing salience based spatial-temporal information preservation, we encode the context awareness as two constraints, which aim at encouraging uniform spatial-temporal transformation across same object. Finally, minimizing the total objective function is formulated to solve a quadratic programming problem.

In our solution, instead of a divide-and-rule strategy, we adopt a context-aware manner to preserving the spatial geometry and temporal coherence. The main contribution of this paper are summarized here.

- A grid graph model is proposed for grid-cell-wise motion layer partition and measurement of contextual relationship. Compared with optical flow, this method produces more robust and reliable estimation. Most importantly, with the inherent concern of the spatial-temporal correlation, the resulting motion layer not only can deliver the temporal motion information but also include the information of contextual relationship implicitly, which plays a critical role for achieving context-aware retargeting.
- We encode the context awareness as two novel constraints during the process of optimization. These two energy penalty terms guiding the regions of same object undergo consistent spatial and temporal transformation, respectively, which refrain retargeted video from obvious spatial-temporal artifacts.
- The aligned grids and grid-cell-wise motion estimation greatly reduce the amount of variables involved in the process of optimization, which make our system work in a highly efficient manner.

II. Related Work

A. Image Retargeting

Many content-aware image retargeting approaches have been proposed, such as cropping [1]–[4], seam carving [5]–[8], warping [9]–[11], and hybrid approaches [12]–[14].

Cropping-based approaches [1]–[4] search for a window of target aspect ratio which covers the most important contents and take this window as the output while completely discarding the part outside.

Seam-based approaches [5] search for an optimal seam which actually is a continuous chain of the pixels from each row or column with the least importance and resize an image by reducing or adding seams iteratively. Several notable works are proposed for improving the original Seam Carving, which can be found in [6]–[8].

Warping-based approaches attempt to transform images continuously. Deformation is more or less allowed in unimportant regions, while the geometry is retrained well for important regions. There are one-directional image warping [9] and omni-directional image warping methods [10]. Li *et al.* propose a dynamic grid partition strategy to preserve the important region precisely [11]. Panozzo *et al.* in [15] use axis-aligned deformation grids for image retargeting.

For hybrid approaches, Rubinstein *et al.* [12] propose a multi-operator strategy, where seam carving is combined with cropping and scaling. This approach improves the retargeting quality through maintaining the spatial structure of whole image. In [13], Sun *et al.* combines a discrete seam-carving algorithm with continuous warping for thumbnail browsing. In [14], seam carving and scaling are used in combination to preserve important regions and global visual effects. Liu *et al.* [16] define a compound operator of cropping and warping to improve the aesthetic composition of images.

Some other works [17]–[19] use coarser patches instead of pixels to quantify the coherence between the original image and target image. The shift-map methods [20], [21] are to optimize the cropping and blending of the important image regions to construct the target image. Pritch *et al.* [20] achieve a better result at the cost of significant change on the image structure. Since the nonuniform salience map usually influences the performance of imaging obviously; in [21], an importance filter is employed for constructing a structure-consistent importance map in order to improve the shift-map.

B. Video Retargeting

Compared with image retargeting, video retargeting is more challenging, since the target video is required to retain not only the aspect ratio and completeness of important object spatially, but also the object motion temporally.

The cCropping strategy is introduced to resolve the video problem in [22]–[25]. The primary strategy could be described as searching for an optimal series of cropping windows to perform a smooth virtual camera motion and presenting the most salient content simultaneously.

Rubinstein *et al.* extend the seam approach for video retargeting in [7]. Through graph-cut, they iteratively carve the 3-D manifolds out from a video cube to adapt the video-to-target aspect ratio. In [8], Grundmann *et al.* relax the constraints to the spatial-temporal structure of seams for motion-aware video retargeting.

In [26]–[28], the warping-based is are proposed for video retargeting, where the preservation of temporal coherence is embodied by constraining regions from temporally inconsistent deformation. Since the motions in video are not given sufficient consideration, these methods usually produce unsatisfactory results when the video contains complex motions. Wang *et al.* in [29] calculate the camera motion and foreground motion, respectively, to ensure the consistent resizing of same objects. In addition, Greisen *et al.* in [30] use axis-aligned deformation grids for video retargeting.

Wang *et al.* in [31], [32] further propose the hybrid approaches, where the cropping is combined with warping. However, the spatial-temporal coherence of local regions in video is preserved independently of each other, which usually

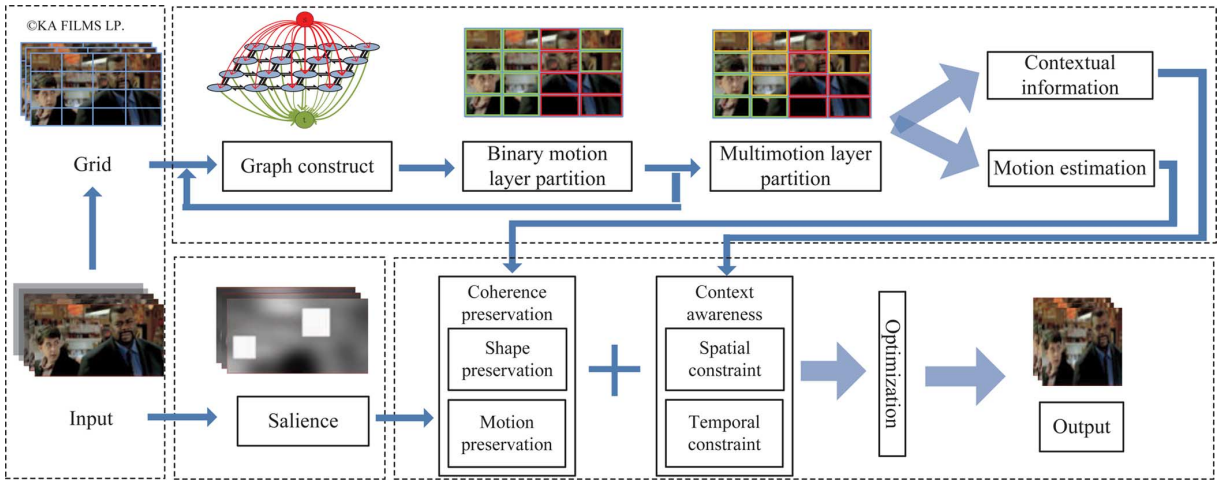


Fig. 1. Pipeline of our algorithm, the graph model guides the optimization both temporally and spatially.

causes the inevitable artifacts both in spatial domain and temporal domain.

In addition to the multimedia techniques, some related Internet techniques are developed to jointly promote the online mobile video services. The former is to maximize the information delivered from the terminal displaying platform to users, while the latter is to maximize the information delivered from server to client in network. The related network techniques aim at transmitting video from the server to diverse user terminals with the best possible quality, and the work discussed in [33] and [34] solves this through coupling the scalable video coding with multipath routing techniques. These techniques mostly rely on the different available paths, the corresponding bandwidth, and the terminals hardware ability while not taking the content of video itself into account.

III. Overview

For video data, there exist some correlations on both spatial and temporal domains according to the content. The regions belonging to the same object are usually visually similar and have continuous temporal motion. Therefore, our context-aware video retargeting framework is based on this kind of spatial-temporal correlation. Fig. 1 shows the basic pipeline of our retargeting approach. A graph model is proposed instead of a flow method, which could produce the simultaneous measurement of context relationship as well as motion estimation. Then, the context awareness is embodied by two constraints during the process of optimization.

First of all, each frame of input sequence is partitioned into uniform grid cells. The objective function is formulated in terms of the coordinates of grid vertices, where the spatial-temporal deformation of regions are calculated at the unit of grid cell. For each frame, the graph model is constructed to estimate the grid-cell-wise motion. Then, the graph-cut algorithm is executed to partition the grid into two motion layers, and the multilayer partition is obtained through iterations. Due to the inherent concern of spatial-temporal correlation, the resulting motion layers deliver not only the temporal motion information but also the information of contextual relationship implicitly. Next, the objective function of optimization is formulated as two parts. On one hand, according to the salience map and motion estimation, the spatial shape and temporal motion of grid cells are preserved differently. On the other hand, the contextual information is encoded into two additional constraints, which encourages the grid cells of the same object to undergo consistent transformation in both space and time domains. Minimizing the objective function leads to the new coordinates of grid vertices, and the output video is rendered by texture mapping.

IV. Graph-Based Motion-Layer Partition

In [35] and [36], Boykov *et al.* introduced graph into the vision category for energy minimization. Wang *et al.* [37] employed a graph cut for moving object segmentation. Here, we expand the graph model for grid-cell-wise motion estimation. Since the graph model inherently raises the contribution of visual similarity between neighboring grid cells to motion estimation in an explicit manner, our approach could suppress the disturbance of noise to some extent and produce more robust and reliable motion estimation than pixel-wise optical flow. More importantly, due to the correspondence with spatial-temporal correlation, the resulting motion layer includes the information of contextual relationship implicitly, which can be revealed as follows: each connected motion layer usually approximates to a single object.

A. Grid Graph Construction

First, one single frame is uniformly partitioned into $m \times n$ grid cells. Denoting the set of grid cells in the lth frame as $Q^l = \{q_1^l, q_2^l, \ldots, q_{m\times n}^l\}$, the set of vertex coordinates of q_i^l is defined as $P_i^l = \{p_{(i,1)}^l, p_{(i,2)}^l, p_{(i,3)}^l, p_{(i,4)}^l, \} \subseteq R^2$, and the centroid coordinate p_i^l can be calculated as the linear combination of vertex coordinates. Then, a grid graph model $G = \langle V, E\rangle$ is constructed to represent this frame, which consists of a set of nodes V and a set of undirected weighted edges E. All of the grid cells in the current frame are represented by the regular nodes. Source node s represents the current motion layer candidate, and sink node t represents the original motion layer of each

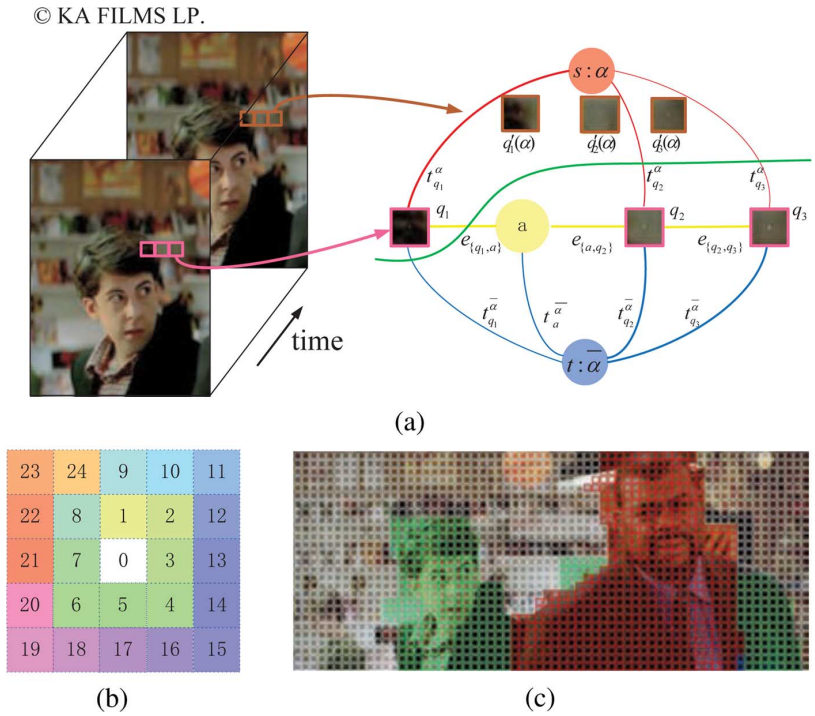


Fig. 2. (a) Construction of the graph model given two temporally neighboring frames, where three horizontally neighboring grid cells are used for a 2-D illustration only. (b) Potential motion trend of each grid cell, which are symbolized as numbers and shown in different colors, for example, $f_{q_{ij}^l} = 18$ means that the content in q_{ij}^l will moves to the position of q_{i+2j-1}^{l+1} in next frame. (c) Resulting motion layer partition.

TABLE I
WEIGHTS OF EDGES CONNECTING THE NODES IN GRAPH

edge	weight	edge	weight
$t_{q_1}^{\alpha}$	$dist_h(q_1, q_1'(\alpha))$	$t_{q_3}^{\bar{\alpha}}$	$dist_h(q_3, q_3'(f_{q_3}))$
$t_{q_2}^{\alpha}$	$dist_h(q_2, q_2'(\alpha))$	$t_a^{\bar{\alpha}}$	$\rho(1 - dist_h(q_1, q_2))$
$t_{q_3}^{\alpha}$	$dist_h(q_3, q_3'(\alpha))$	$e_{\{q_1,a\}}$	
$t_{q_1}^{\bar{\alpha}}$	$dist_h(q_1, q_1'(f_{q_1}))$	$e_{\{a,q_2\}}$	
$t_{q_2}^{\bar{\alpha}}$	$dist_h(q_2, q_2'(f_{q_2}))$	$e_{\{q_2,q_3\}}$	$\rho(1 - dist_h(q_2, q_3))$

grid cell. The t edge linking regular node to terminal node measures the accuracy that the grid represented by the former has the motion represented by the latter; while the e edge between regular nodes reflects the visual similarity of the neighboring grids. As shown in Fig. 2(a), three horizontally neighboring grid cells q_1, q_2, q_3 is used for a 2-D illustration only. In practice, all grid cells will be included in the graph, where the vertically neighboring grids are treated in same manner. We denote the original motion labels of q_1, q_2, q_3 as $f_{q_1}, f_{q_2}, f_{q_3}$. Given motion layer candidate α, we define $q'(\alpha)$ as the corresponding grid cell, in next frame, of q, and the weights of edges are defined as in Table I. The graph-cut algorithm introduced below would decide whether the grid cells are partitioned to new motion layer s or stay the same.

In Table I, $dist_h(q_1, q_2)$ means *Bhattacharyya* distance between histogram between q_1 and q_2, which is used to measure the similarity of them. ρ is a coefficient for adjusting the contribution of visual similarity of neighboring grid cells to motion layer partition. a is an auxiliary node, which is constructed to preserve the min-cut energy in each iteration.

B. Graph Cut

When the grid graph model is constructed, a max-flow/min-cut algorithm [35], [36] is executed for searching an optimal cut, i.e., the green line in Fig. 2(a), and, consequently, all of the grid cells are partitioned into different motion layers. To be more exact, the max-flow/min-cut algorithm is applied to minimizing the energy function as follows:

$$E(F) = E_{data} + \rho E_{smooth} \tag{1}$$

$$E_{data}(F) = \sum_{i,j}^{m,n} dist_h(q_{ij}, q_{ij}'(f)) \tag{2}$$

$$E_{smooth}(F) = \sum_{i,j}^{m,n} |\partial f/\partial i| \cdot (1 - dist_h(q_{ij}, q_{i+1j})) + |\partial f/\partial j| \cdot (1 - dist_h(q_{ij}, q_{ij+1})) \tag{3}$$

where $q_{ij}'(f)$ is the corresponding grid cell, in the next frame, of q_{ij} according to the motion label f. F represents the current motion layer partition on all of the grid cells, i.e., $F = \{f_1, f_2, \ldots, f_{m\times n}\}$. Minimizing equation (1) is to find the optimal F. Equation (2) is responsible for providing the grid cells the motion estimation as accurate as possible at a grid-cell level and corresponds to the t edges in graph. The left-hand terms of (3) are used to finally produce a smooth layer partition; the right-hand terms further encourage the visually similar and neighboring grid cells to be partitioned into the same layer with special emphasis and correspond to the e edge. For the sake of the

huge difference between the grid-cell-level motions, the definition of (3) can be further simplified as follows:

$$E_{smooth}(F) = \sum_{i,j}^{m,n} \mathsf{f}_i \cdot (1 - dist_h(q_{ij}, q_{i+1j})) + \mathsf{f}_j \cdot (1 - dist_h(q_{ij}, q_{ij+1})) \quad (4)$$

$$\mathsf{f}_i = \begin{cases} 1 & if\ f_{ij} \neq f_{i+1j} \\ 0 & if\ f_{ij} = f_{i+1j} \end{cases} \quad \mathsf{f}_j = \begin{cases} 1 & if\ f_{ij} \neq f_{ij+1} \\ 0 & if\ f_{ij} = f_{ij+1} \end{cases} \quad (5)$$

which completely corresponds to the graph structure defined by Fig. 2(a) and Table I.

In this solution, 25 potential motion layers are deemed sufficient to estimate the grid-cell-wise motions, which are shown in Fig. 2(b). Since the graph-cut is a binary partition process, for multilayer partition, the graph-cut algorithm is executed 25 times for one iteration in our configuration. As the proposed motion model needs to execute a graph-cut many times, the min-cut energy obtained by previous graph-cuts should be preserved in the current graph structure to ensure that the resulting motion partition achieve global energy minimization. For this purpose, the auxiliary node a is introduced. Once two neighboring grids originally in same layer are partitioned into different layers by the current cut, in the next graph structure, an auxiliary node will be inserted between them. This node and the related edges ensure that for the next graph-cut process, if no grid cell is partitioned into the new layer, the corresponding min-cut will pass through only the edges linking to the sink node t and keep the equal energy to the previous one.

In addition, since the motion between consecutive frames may be too small to be captured by our grid-cell-wise approach, we estimate the motion between frames at some regular interval. According to our experiments, $2 \sim 4$-frame interval is appropriate in most cases.

C. Measurement of Contextual Relationship

As shown in Fig. 3, the proposed motion model is compared with the state-of-the-art optical flow technique [38]. Optical flow achieves the better pixel-wise accuracy in estimation, while is weak in reflecting the wholeness of objects. This is because the smooth process of optical flow considers the influence of all of the neighboring pixels indiscriminately and lacks the consideration on spatial information. As shown in Fig. 3, the flow estimation on the distant part of "bridge" is more influenced by the background and become undistinguishable. On the other hand, although the grid graph model sacrifices the pixel-wise accuracy, this loss actually exerts limited negative influence to the quality of retargeting. More importantly, the grid graph model raises the contribution of visual similarity of neighboring regions to motion estimation in an explicit manner and tends to partition the visually similar grids into the same layer. As a result, the resulting motion layers keep the wholeness of moving objects better and include information of contextual relationship implicitly. As shown in Fig. 3, the resulting motion layers describe the left "character" and right "bridge" more completely. In addition, the grid graph model provides a more efficient motion estimation with an average conversion time of 8.6 s compared with 51.7 s of optical flow.

Optical flow Motion layer partition

Fig. 3. Comparison of motion layer partition and optical flow. The flow algorithm process the frames with the same interval as the proposed approach.

According to the optimal motion partition of a grid cell denoted as $f_{q_1}^{\text{op}}$, $f_{q_1}^{\text{op}}$, and $f_{q_1}^{\text{op}}$, we define a set $B = \{\ldots, b_{q_1q_2}, b_{q_2q_3}, \ldots\}$ to present the contextual relationship between neighboring grid cells approximately. For the neighboring grid cells q_1, q_2, $b_{q_1q_2}$ is defined as follows:

$$b_{q_1q_2} = \begin{cases} 1, & \text{if} \quad f_{q_1}^{op} = f_{q_2}^{op} \\ 0, & \text{if} \quad f_{q_1}^{op} \neq f_{q_2}^{\text{op}}. \end{cases} \quad (6)$$

Equation (6) quantifies the spatial-temporal correlation between neighboring regions binarily, which factually reflects the possibility that grid cells q_1 and q_2 belong to the same object. This measurement of a contextual relationship plays an critical role for smooth and spatial-temporal consistent retargeting.

V. Context-Aware Grid Optimization

Here, we describe the optimization procedure of our grid-based retargeting framework. Due to the motion estimation and measure of context relationship obtained in Section IV, the total objective function incorporates two additional constraints as well as spatial-temporal coherence preservation in order to achieve the context aware retargeting. The objective function is formulated in terms of the coordinates of grid vertices. Minimizing the energy function under some constraints results in the new vertex positions, and the output video is rendered through texture mapping.

A. Salience-Based Spatial-Temporal Coherence Preservation

Spatial Shape Preservation: As discussed above, the energy function is formulated in terms of the deformed vertex coordinates. We employ the centroid coordinate p_i to present the position of grid cell q_i, which can be calculated according to the four vertex coordinates directly. In our solution, the grid is further requested to be axis aligned, i.e., the grid cells in the same row (or column) have the same height (or width). This aligned structure decreases the amount of variables and reduces the computation complexity dramatically, which provides a convenient way for multiframe video retargeting. We formulate the spatial deformation energy of a single frame with $m + 1 + n + 1$ variables as follows, which reaches up to $2 \times (m + 1) \times (n + 1)$ in a non-aligned grid framework [10] as follows:

$$D_s^l = \sum_{i}^{m \times n} \left(\tilde{h}_i^l - ars \cdot \tilde{w}_i^l \right)^2 \cdot s_i^l \quad (7)$$

where D_s is the deformation accumulation of all $m \times n$ grid cells in the lth frame, $\tilde{q}_i$ represents the output grid, and $\tilde{w}_i$ and $\tilde{h}_i$ represent the width and height of $\tilde{q}_i$, respectively, and s_i is defined as the visual importance of grid cell q_i, which ranges

from 0.2 to 1. The lower limit of s_i is set to be nonzero so as to prevent the deformed grids from undue distortion and the incorrect overlap between grids; ars is the aspect ratio of original grid cell.

Equation (7) is supposed to preserve spatially the aspect ratio of important regions in the targeted video.

Temporal Motion Preservation: Beside preserving the spatial content, the temporal motion should be maintained consistent with the source video. In Section IV, the grid graph model has estimated the motions of grid cells in each frame. Given the motion labels $f_{q_i^l}$ and $f_{q_j^l}$, the corresponding grid cells of q_i^l and q_j^l in the $l+1$th frame are represented as q_u^{l+1} and q_v^{l+1}. We formulate an energy term D_t^l to preserve the motions of grid cells, especially for the salient ones and define it as follows:

$$D_t^l = \sum_i \left\| \left(\tilde{p}_u^{l+1} - \tilde{p}_i^l\right) - kp \cdot \left(p_u^{l+1} - p_i^l\right) \right\|_2^2 \cdot s_i^l \quad (8)$$

where $\tilde{p}_i$ represents the centroid point of $\tilde{q}_i$, and $kp = [kpx, kpy]^T$ is a scale factor controlling the motion difference between the grid cells in output and those in source. In practice, we make kp work in a heuristic way: in retargeted video, we believe the changes of object movements in horizontal direction and vertical direction should be proportional to the width change and height change of video, i.e., $kpx = W_T/W_S$, $kpy = H_T/H_S$, where the resolution of the source and target video are $W_S \times H_S$ and $W_T \times H_T$, respectively. Equation (8) ensures the salient regions have consistent motion trends with in source video.

B. Context-Aware Spatial-Temporal Constraints

1) Spatial Constraints: According to the salience map, we preserve the spatial coherence at the unit of grid cell. Unfortunately, the importance map is not always consistent with the spatial structure of original image such that the map usually changes a great deal for the same object. This variety results inevitably in the nonuniform deformation of the object, which leads to the obvious damage to geometrical property. Some researchers have noticed this problem and have begun to solve it. In [21], Ding *et al.* employ an importance filter for redistributing the salience map on the same object to be homogeneous. Unlike them, we suppress the nonuniform deformation of object via a spatial context constraint.

As discussed in Section IV, the resulting motion-layer partition implicitly includes the information of contextual relationship, and each connected motion layer usually approximates to one single object. Here, we formulate a spatial context constraint as a penalty term of energy function

$$P_s^l = \sum_{b_{ij}^l \in B} \mathrm{diff}_s\left(\tilde{q}_i^l, \tilde{q}_j^l\right) \cdot b_{ij}^l \quad (9)$$

$$\mathrm{diff}_s\left(\tilde{q}_i^l, \tilde{q}_j^l\right) = \left(\tilde{w}_i^l - \tilde{w}_j^l\right)^2 + \left(\tilde{h}_i^l - \tilde{h}_j^l\right)^2 \quad (10)$$

where $B = \{\ldots, b_{ij}^l, \ldots\}$ is the set defined in Section IV, which can be considered as the quantification of spatial-temporal correlation between neighboring grid cells. The element b_{ij}^l approximately reflects the possibility that the neighboring grid cells q_i^l and q_j^l belong to the same object, $diff_s(\cdot,\cdot)$ evaluates the difference of spatial structure between neighboring grid cells, and w and h are the width and height of the corresponding grid cell, respectively. Equation (9) demonstrates that the regions of the same object undergo consistent spatial deformation.

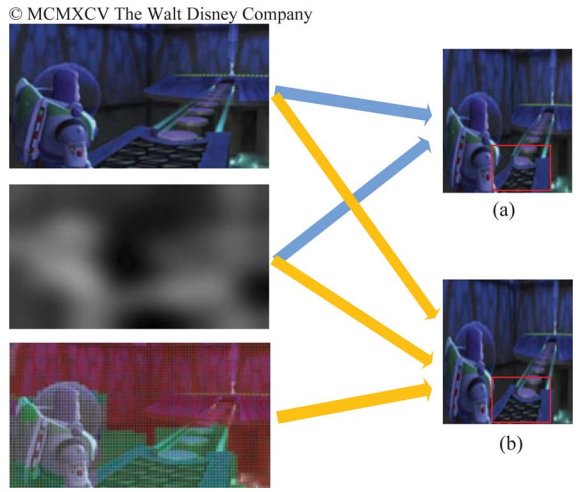


Fig. 4. Example of video retargeting with spatial constraint. The first column shows the source frame, the salience map, and the motion layers labeled by graph model, respectively. (a) Resizing through preserving the content of each grid cell independently. (b) Resizing by taking the spatial constraint into account.

Fig. 4 demonstrates the effectiveness of this kind of spatial constraint. In Fig. 4(a), the source frame is resized through deforming each grid cell independently according to the salience map. Note that, because of the nonhomogeneous salience map, the bridge is distorted significantly, while our algorithm partitions most of the grid cells lapping over the bridge into the same layer by the graph model and consequently makes them maintain a similar aspect ratio, as shown in Fig. 4(a). As result, the retargeted video avoids the spatial distortion effectively, as shown in Fig. 4(b).

2) Temporal Constraints: The lack of consideration of contextual relationship leads to not only spatial distortion but also temporal artifacts, such as the foreground jitter and background waving. This is because the regional motions of the same object are generally continuous, while the respective manner of preserving temporal coherence of regions may make them undergo nonuniform temporal transformation and damage the continuity. For example, the left column of Fig. 5 shows two temporally neighboring source frames and the corresponding motion-layer partition. The retargeting results of MVR [31] are shown in the middle column, which neglects the contextual relationship between grid cells. We track two points on the girl's face and label them in red and orange, respectively. Note that the motion vectors of two points differ from each other both in amplitude and direction obviously, which consequently allows the face to go through temporally inconsistent deformation.

To achieve the context-aware temporal coherence preservation, we encode the temporal context constraint by the set B in Section IV as follows:

$$P_t^l = \sum_{b_{ij}^l \in B} \mathrm{diff}_t\left(\tilde{q}_i^l, \tilde{q}_j^l\right) \cdot b_{ij}^l \quad (11)$$

$$diff_t\left(\tilde{q}_i^l, \tilde{q}_j^l\right) = \left\| \left(\tilde{p}_u^{l+1} - \tilde{p}_i^l\right) - \left(\tilde{p}_v^{l+1} - \tilde{p}_j^l\right) \right\|_2^2 \quad (12)$$

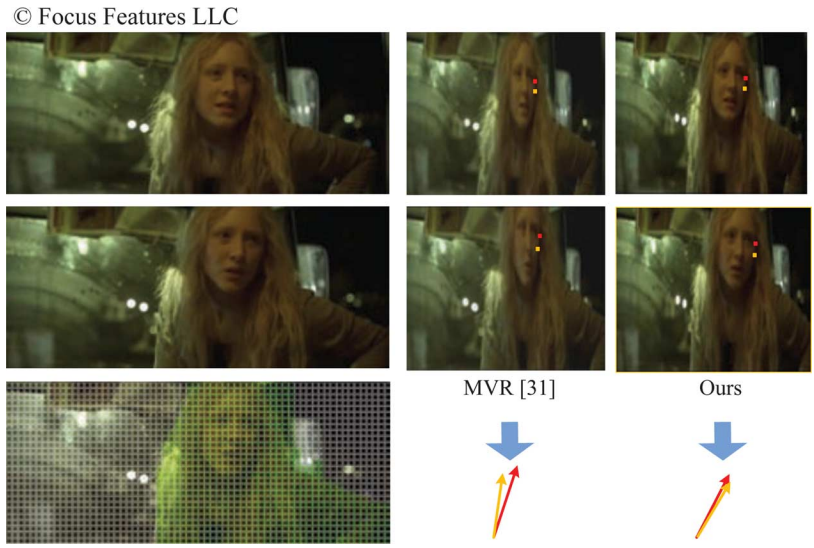


Fig. 5. Example of video retargeting with temporal constraint. The first column includes two temporally neighboring frames and the corresponding motion-layer partition. The second and third columns show the retargeted frames by MVR [31] and our approach, respectively, and our approach could make the motion vectors of two chosen points more consistent.

where $diff_t(q_i^l, q_j^l)$ evaluates the difference of motion of q_i^l and q_j^l. Equation (11) assures that the regional motions of the same object undergo consistent temporal transformation. In other words, the energy term equation (11) develops the grid-cell-level temporal consistency achieved through (8) into the object-level temporal consistency. In addition, some existing work includes the previous grid deformation as an additional temporal constraint, and the combination of (8) and (11) factually does the same job in this method. The combined effect of both temporal energy terms plays a critical role for the proposed context-aware retargeting solution.

Fig. 5 demonstrates the effectiveness of temporal constraint. The right column shows our retargeting results, where the motion vectors of two chosen points are basically consistent. Compared with other methods, our approach achieves the temporal smoothness and avoids the noticeable temporal artifacts. More proof can be found in our supplementary material.

C. Energy Minimization

By combing spatial energy and temporal energy together, the total objective function is finally formulated as follows:

$$D = \sum_l (D_s + \lambda P_s + D_t + \delta P_t) \tag{13}$$

where λ and δ are the weights of spatial and temporal constraints, respectively, and our approach works well when $\lambda = 0.4$ and $\delta = 10$. When $\lambda = 0.4$, the spatial context constraint is sufficiently powerful to repair the nonuniform deformation occurring on same object, which results from the nonhomogeneous salience map, while not unduly influencing the flexibility of distributing distortion. The value of delta is relatively high due to the following fact: in the process of retargeting, the motion difference of the neighboring grids belonging to the same layer is usually very small. Hence, in relation to the other energy terms, a high δ is expected to make sure the temporal context constraint play its due role.

TABLE II
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART TECHNIQUES. THE RESULTING TIME MEASURES ONLY OPTIMIZATION WHILE NOT INCLUDING THE SALIENCE MAP, FACE DETECTION, AND MOTION ESTIMATION. THREE ALGORITHMS HANDLE THE VIDEO SEQUENCES OF ABOUT 700 × 300 RESOLUTION AND 200-FRAMES LENGTH WITH A GRID CELL OF 10 ×,10

Method	MVR[31]	SCVR[32]	Ours
Time	63s	10s	9s
Hardware	2.66GHz CPU 8GB RAM		2.9GHz CPU 4GB RAM

The total objective function is a quadratic function in terms of the vertex coordinates of grids. Linear equality constraints are imposed to make sure that each deformed frame has targeted width and height. Thus, the optimization is to solve a convex quadratic programming, which can be achieved through numerical method. In this solution, a CPU-based conjugate gradient solver is used to minimize the objective function, and the optimization is initialized with the uniform grids satisfying all constraints. There are N unknowns and four equality constraints for each frame and L frames involved in optimization, and the computation complexity is $O(4L \cdot L \cdot N)$.

VI. EXPERIMENTAL RESULTS

We test our algorithm by retargeting videos to 50% of the original width. At first, a shot detection algorithm [39] is employed, and each shot is retargeted individually. For the salience map, we apply the visual attention-based method [40] together with the face detector [41]. For comparison, we run our approach on the dataset provided in [32], which includes a variety of videos in order to reflect the universality. This data set ranges from scenes with relatively fixed foreground to scenes containing noticeable object movement, from scenes containing single moving object to scenes containing multiple moving objects, from scenes captured with no camera motion to scenes captured with typical camera motion (e.g., pan or scan). The experimental materials are with the spatial resolution of 700 × 300

Fig. 6. Failing example.

and temporal resolution of 30 fps, and the length ranges from 80 to 230 frames.

In our solution, both the spatial and temporal deformations are calculated at the grid cell level, and the frame interval relates to the grid size. A fine grid leads to a high quality of output but introduces a smaller frame interval, more variables in optimization, and higher computation time consequently. Based on practical experience, we found the grid cell size of 8×8 and motion estimation of a two-frame interval, together with $\lambda = 0.4$ and $\delta = 10$, strike a good balance between the imaging quality and operation efficiency in most cases.

A. Performance

Benefitting from the aligned grid structure and grid-cell-wise motion estimation, the proposed system could work in a high-efficiency manner. In our experimental setting, the whole video sequence of one shot is used as the input for the globally optimal spatial-temporal effectiveness. Table II shows the time cost comparison with the state-of-the-art grid-based techniques on handling the video sequences of similar resolution and length with same grid size. The proposed approach achieves a compatible performance. Note that SCVR, which achieves the optimal performance, benefits from the parallel processing comparing with our single thread process. On the other hand, the formulated objective function, relying on the current frame and the next frame only, makes it possible to process videos scalably, even in real time on incoming video. For the video stream of the same resolution, the computation time is approximately 11 ms per frame.

B. Quality

Our approach is compared with state-of-the-art methods including stream video retargeting (SVR) [28], motion-based video retargeting (MVR) [31], and scalable and coherent video retargeting (SCVR) [32]. Some of experimental results are shown in Fig. 7. In general, our method achieves the satisfactory temporal-spatial coherent retargeting while avoiding various artifacts and shows obvious superiority over other methods.

For SVR [28], the content lying in the middle of frame is retained preferably. When there is no complex foreground motion, SVR can produce compatible results with ours. However, when the video contains complex motion, the abrupt artifacts usually appear near the edge, e.g., the car body in the first row, the unnatural shoulder of the black man in the second row, and the distortion of window frame in the sixth row in Fig. 7 This can be explained as lack of motion awareness. In addition, the girl in the fifth row and the cartoon in the ninth row suffer more distortion compared with our method. MVR [31], taking the motion information into account explicitly, usually provides a satisfactory performance. However, the noticeable artifacts still happen. As shown in Fig. 7, serious distortion occurs on the foreground in the second, fourth, fifth, seventh, and eighth rows. The background waving occurs in the third row; the cartoon is down-scaled a lot in the ninth row. Since the cropping strategy has been incorporated to improve the quality of retargeting, there are always some content discarded directly. SCVR [32] achieves a great balance between retaining prominent content and preserving temporal coherence. Similar to MVR, SCVR also incorporates cropping to free up more space for preserving salient content. Although SCVR usually brings viewers comfortable watching experience, inevitable loss of visual information is incurred at the same time, e.g., the building next to the street in the first row, the window frames in the sixth row, and the excavator in the seventh row in Fig. 7.

C. User Study

In order to evaluate our approach objectively, we perform a user study. There are altogether 18 video sequences and 46 volunteers involved, and SVR [28], MVR [31], SCVR [32], and our approach are compared in this study. Each pair of approaches was tested 4968 times in all. Thirty-four participators have the higher education background, and 18 of them come from natural science realm while the remaining 16 come from the social

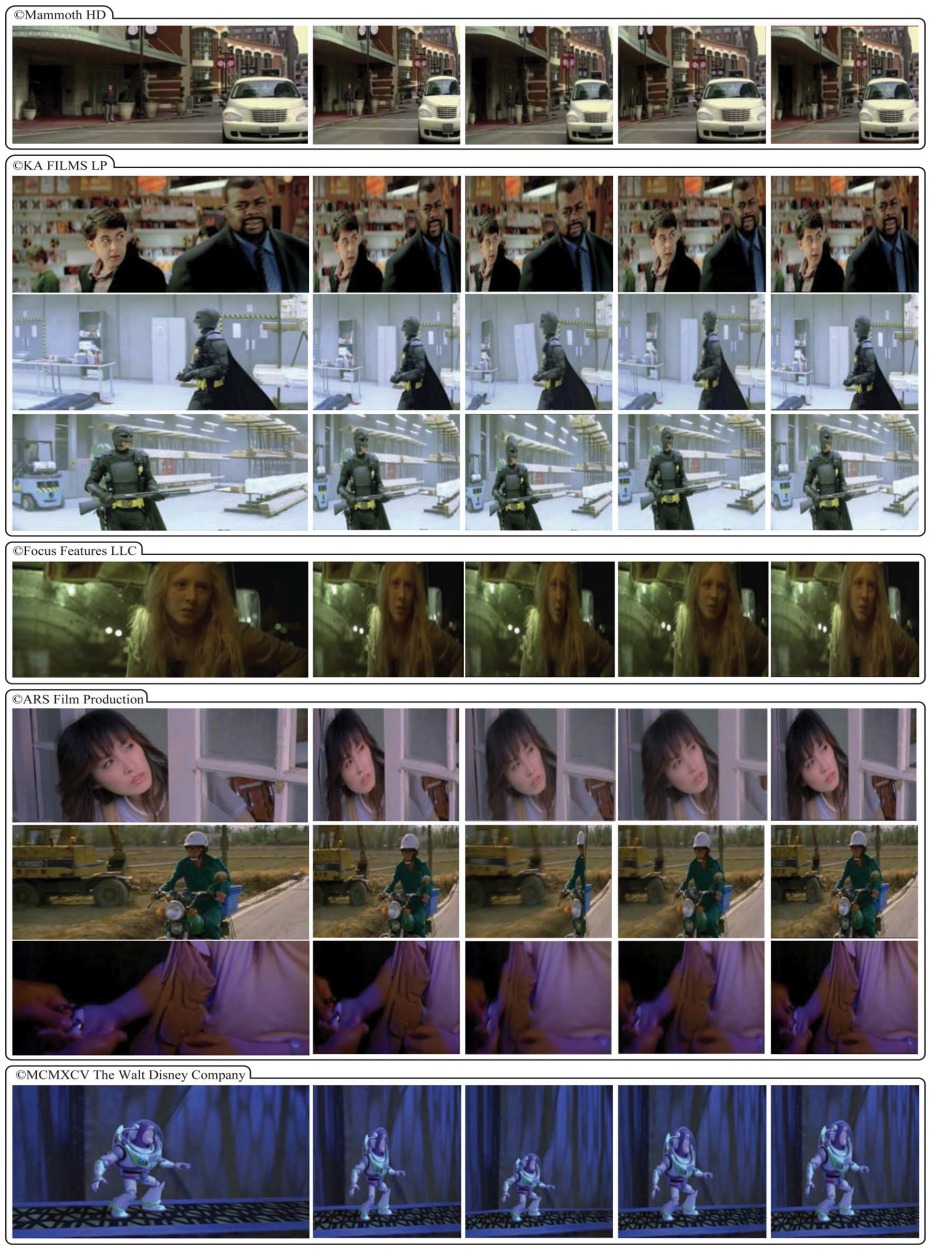


Fig. 7. Experimental results. (a) Original. (b) SVR [28]. (c) MVR [31]. (d) SCVR [32]. (e) Ours.

science realm. In addition, 12 participants are teenagers, who are usually more familiar with dynamic photograph means and are aesthetically more sensitive to the videos containing complex movements. Different methods are compared in pairs at a random order. Not influencing the participants' aesthetic concept, no hints or *a priori* knowledge are delivered to them in advance. The participators are asked to vote their preference between the two versions of retargeting.

The statistics of all of the feedback are shown in Table III. Comparing with SVR, our method is preferred at the rate 71.7%. It results fthat our method considers the motions in video more directly. Compared with MVR, our method shows more popularity with the preference of 73.9% participators. This is because our method suppresses the distortion of foreground in output effectively. Compared with SCVR, the percentage becomes 63%, where our approach shows the superiority on avoiding temporal artifacts. In the comparisons, the overall percentage of preference to our method is 69.5%, which demonstrates that our method can produce more acceptable retargeting results in contrast to 39.2% of SVR, 33.3% of MVR, and 58% of SCVR.

D. Failing Case

Although our approach takes the context awareness into account, visual artifacts still occur occasionally. A failing example is shown in Fig. 6. Note that the bottom of the frame retargeted by our approach is vertically stretched by mistake, which damages the structure of the tree and leads to an unfavorable comparison with other retargeting methods. In essence, this artifact comes from the rectangularity of grid cell. Compared with the

TABLE III
STATISTICS OF USER STUDY. THE TABLE PRESENTS THE PERCENTAGE OF PREFERENCE TO EACH APPROACH

% of preference to →	Ours	SVR	MVR	SCVR	Mean
Ours	-	71.7	73.9	63	69.5
SVR [28]	28.3	-	56.6	32.6	39.2
MAR [31]	26.1	43.4	-	30.4	33.3
SCVR [32]	37	67.4	69.6	-	58

unconfined grid used by MVR [31], rectangular grid cells simplify the computation for video issue significantly but sacrifice the flexibility in redistributing the spatial distortion. In Fig. 6, the average importance of the bottom of the frame is relatively lower than other regions. As result, the grid cells in the bottom row have to suffer more vertical stretch.

Moreover, when the salient regions are distributed across the whole frame uniformly, our approach may not produce a better result than cropping incorporated methods such as MVR. Compared with our approach, MVR factually frees up more space for absorbing the spatial distortion by directly cropping out the leftmost and rightmost columns. This can be reflected by the example shown in Fig. 6.

VII. CONCLUSION

In this paper, we propose a context-aware framework instead of the divide-and-rule strategy for video retargeting. The graph model is employed for modeling the contextual relationship between grid cells binary as well as the grid-cell-wise motion estimation. The context awareness is embodied through encoding the measurement of contextual relationship as two additional constraints in the space and time domains, respectively, during the optimization. Compared with state-of-the-art techniques, our approach refrains the regions of the same object from the inconsistent spatial-temporal transformation and shows obvious superiority in suppressing spatial-temporal artifacts. Beyond that, the high efficiency achieved by the aligned grid and grid-cell-wise motion estimation promises that our approach is a good practical prospect.

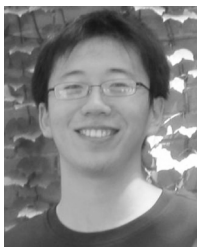
ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for the constructive comments and valuable suggestion. The authors would also like to thank Y.-S. Wang, P. Krähenbühl, and M. Lang for providing the retargeted video clips for comparison.

REFERENCES

[1] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *ACM Multimedia Syst. J.*, vol. 9, no. 4, 2003.
[2] H. Liu, X. Xie, W.-Y. Ma, and H.-J. Zhang, "Automatic browsing of large pictures on mobile devices," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 148–155.
[3] B. Suh, H. Ling, B.-B. Bederson, and D.-W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proc. UIST*, 2003, pp. 95–104.
[4] A. Santella, M. Agrawala, D. Decarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semiautomatic photo cropping," in *Proc. CHI*, 2006, pp. 771–780.
[5] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graphics*, vol. 26, no. 3, 2007.
[6] A. Mansfield, P. Gehler, L. V. Gool, and C. Rother, "Scene carving: Scene consistent image retargeting," in *Proc. ECCV*, 2010.
[7] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graphics*, vol. 27, no. 3, 2008.
[8] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Discontinuous seam-carving for video retargeting," in *Proc. CVPR*, 2010.
[9] R. Gal, O. Sorkine, and D. Cohen-or, "Feature-aware texturing," in *Proc. EGSR*, 2006, pp. 297–303.
[10] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scale-and-stretch for image resizing," *ACM Trans. Graphics*, vol. 27, no. 5, 2008.
[11] B. Li, Y.-M. Chen, J.-Q. Wang, L.-Y. Duan, and W. Gao, "Fast retargeting with adaptive grid optimization," in *Proc. ICME*, 2011, pp. 1–4.
[12] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Trans. Graphics*, vol. 28, no. 3, 2009.
[13] J. Sun and H.-B. Ling, "Scale and object aware image retargeting for thumbnail browsing," in *ICCV*, 2011.
[14] W.-M. Dong, N. Zhou, J.-C. Paul, and X.-P. Zhang, "Optimized image resizing using seam carving and scaling," in *SIGGRAPH*, 2009.
[15] D. Panozzo, O. Weber, and O. Sorkine, "Robust image retargeting via axis-aligned deformation," *Comput. Graphics Forum*, vol. 31, no. 2, pp. 229–236, 2012.
[16] L.-G. Liu, R.-J. Chen, L. Wolf, and D. Cohen-Or, "Optimize photo composition," *Comput. Graphic Forum*, vol. 29, no. 2, 2010.
[17] T.-S. Cho, M. Butman, S. Avidan, and W.-T. Freeman, "The patch transform and its applications to image editing," in *Proc. CVPR*, 2008.
[18] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. CVPR*, 2008.
[19] C. Barnes, E. Shechtman, A. Finkelstein, and D.-B. Gold-Man, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graphics*, vol. 28, no. 3, 2009.
[20] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," in *Proc. ICCV*, 2009, pp. 151–158.
[21] Y.-Y. Ding, J. Xiao, and J.-Y. Yu, "Importance filtering for image retargeting," in *Proc. CVPR*, 2011.
[22] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in *Proc. Multimedia*, 2006, pp. 241–250.
[23] C. Tao, J. Jia, and H. Sun, "Active window oriented dynamic video retargeting," in *Proc. ICCV*, 2007.
[24] T. Deselaers, P. Dreuw, and H. Ney, "Pan, zoom, scan time-coherent, trained automatic video cropping," in *Proc. CVPR*, 2008.
[25] M. L. Gleicher and F. Liu, "Re-cinematography: Improving the camerawork of casual video," *ACM Trans. Multimedia Comput. Commun, Appl.*, vol. 5, no. 1, pp. 1–28, 2008.
[26] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. ICCV*, 2007.
[27] Y.-F. Zhang, S.-M. Hu, and R. R. Martin, "Shrinkability maps for content-aware video resizing," , 2008.
[28] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graphics*, vol. 28, no. 5, 2009.
[29] Y.-S. Wang, H. Fu, O. Sorkine, T.-Y. Lee, and H.-P. Seidel, "Motion-aware temporal coherence for video resizing," *ACM Trans. Graphics*, vol. 28, no. 5, 2009.
[30] P. Greisen, M. Lang, S. Heinzle, and A. Smolic, "Algorithm and VLSI architecture for real-time 1080p60 video retargeting," *High Performance Graphics*, pp. 57–66, 2012.
[31] Y.-S. Wang, H.-C. Lin, O. Sorkine, and L. T.-Y. , "Motion-based video retargeting with optimized crop-and-warp," *ACM Trans. Graphics*, 2010.
[32] Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee, "Scalable and coherent video resizing with per-frame optimization," *ACM Trans. Graphics*, vol. 30, no. 4, 2011.
[33] S. Li, Z. Zhu, W. Li, and H. Li, "Efficient and scalable cloud-assisted SVC video streaming through mesh networks," in *Proc. ICNC*, 2012, pp. 944–948.
[34] M. Ghareeb, A. Ksentini, and C. Viho, "Scalable video coding (SVC) for multipath video streaming over video distribution networks (vdn)," in *Proc. ICOIN*, 2011, pp. 206–211.
[35] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[36] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124, 1137, Sep. 2004.

[37] J. Wang, H.-F. Wang, Q.-S. Liu, and H.-Q. Lu, "Automatic moving object segmentation with accurate boundaries," in *Proc. ACCV*, 2006.
[38] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. CVPR*, 2010, pp. 2432–2439.
[39] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Proc. CVPR*, 2003, vol. 2, no. 2, pp. 343–348.
[40] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. CVPR*, 2007.
[41] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, 2004.



Zhan Qu received the B.E. degree from Inner Mongolia University, China, in 2007, and the M.S. degree from Tianjin University, Tianjin, China, in 2009 . He is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His primary research interests include image/video analysis and processing, mobile multimedia, intelligent video surveillance, pattern recognition, and computer vision.

Jinqiao Wang (M'09) received the B.E. degree from Hebei University of Technology, Tianjin, China, in 2001, the M.S. degree from Tianjin University, Tianjin, China, in 2004, and the Ph.D. degree in pattern recognition and intelligence systems from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently an Assistant Professor with the Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition and machine learning, image and video processing, mobile multimedia, and intelligent video surveillance.

Min Xu (M'12) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2000, the M.S. degree from the National University of Singapore, Singapore, in 2004, and the Ph.D. degree from the University of Newcastle, Newcastle, Australia, in 2010.

Currently, she is a Lecturer with the School of Computing and Communications, Faculty of Engineering, and Information Technology, University of Technology, Sydney, Australia. Her research interests include multimedia content analysis, video adaptation, interactive multimedia, pattern recognition, and computer vision.

Hanqing Lu (SM'06) received the B.E. and M.E. degrees from Harbin Institute of Technology, Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree from Huazhong University of Sciences and Technology, Wuhan, China, in 1992 .

Currently, he is a Deputy Director with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include image and video analysis, medical image processing, and object recognition. He has authored and coauthored more than 100 papers in these fields.