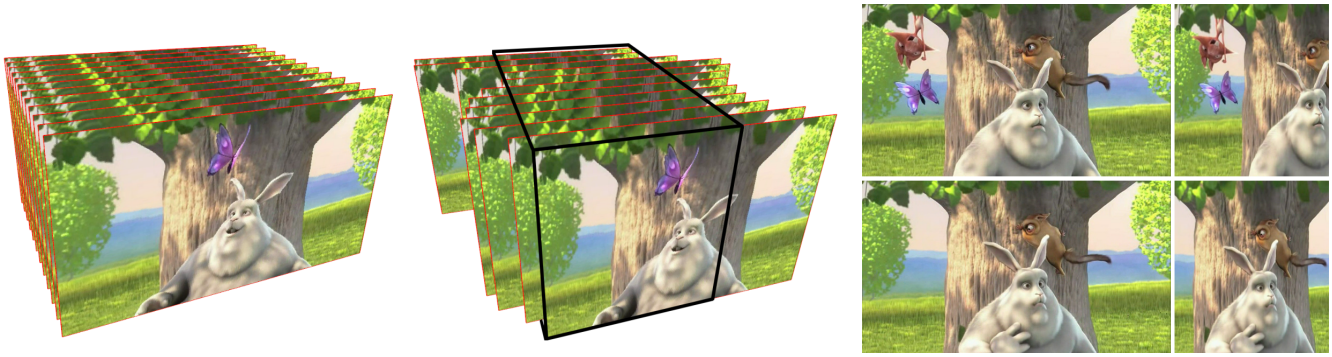
Motion-based Video Retargeting with Optimized Crop-and-Warp

Yu-Shuen Wang[1,2] Hui-Chih Lin[1] Olga Sorkine[2] Tong-Yee Lee[1]

[1]National Cheng-Kung University, Taiwan [2]New York University



©Blender Foundation

Figure 1: *Our video retargeting framework has motion information at its core, utilizing it to define temporal persistence of video contents and to describe temporal coherence constraints. Left: We combine cropping and warping by forcing all informative video content inside the target video cube, without a priori constraining the size of each frame. Right: parts of the bunny and the squirrel are allowed to be cropped (top) since they will fully appear later in the video (bottom).*

Abstract

We introduce a video retargeting method that achieves high-quality resizing to arbitrary aspect ratios for complex videos containing diverse camera and dynamic motions. Previous content-aware retargeting methods mostly concentrated on spatial considerations, attempting to preserve the shape of salient objects in each frame by removing or distorting homogeneous background content. However, sacrificeable space is fundamentally limited in video, since object motion makes foreground and background regions correlated, causing waving and squeezing artifacts. We solve the retargeting problem by explicitly employing motion information and by distributing distortion in both spatial and temporal dimensions. We combine novel cropping and warping operators, where the cropping removes temporally-recurring contents and the warping utilizes available homogeneous regions to mask deformations while preserving motion. Variational optimization allows to find the best balance between the two operations, enabling retargeting of challenging videos with complex motions, numerous prominent objects and arbitrary depth variability. Our method compares favorably with state-of-the-art retargeting systems, as demonstrated in the examples and widely supported by the conducted user study.

Keywords: video retargeting, cropping, warping, spatial and temporal coherence, optimization

ACM Reference Format
Wang, Y., Lin, H., Sorkine, O., Lee, T. 2010. Motion-based Video Retargeting with Optimized Crop-and-Warp. *ACM Trans. Graph. 29*, 4, Article 90 (July 2010), 9 pages. DOI = 10.1145/1778765.1778827 http://doi.acm.org/10.1145/1778765.1778827.

1 Introduction

Retargeting images and video for display on devices with different resolutions and aspect ratios is an important problem for the modern society, where visual information is accessed using a variety of display media with different formats, such as cellular phones, PDAs, widescreen television, and more. To fully utilize the target screen resolution, traditional methods homogeneously rescale or crop the visual content to fit the aspect ratio of the target medium. Simple linear scaling distorts the image content, and cropping may remove valuable visual information close to the frame periphery. To address this problem, content-aware retargeting techniques were recently introduced. These methods non-homogeneously deform images and video to the required dimensions, such that the appearance of visually important content is preserved at the expense of removing or distorting less prominent parts of the input.

Most content-aware retargeting techniques to date have concentrated on *spatial* image information, such as various visual saliency measures and face/object detection, to define visually important parts of the media and to guide the retargeting process. They rely on the fact that removing or distorting homogeneous background content is less noticeable to the eye [Shamir and Sorkine 2009]. Recent video retargeting works [Krähenbühl et al. 2009; Wang et al. 2009] additionally average the per-frame importance maps over several frames and grant higher importance to moving objects to improve the temporal coherence of the result.

Yet, video retargeting is fundamentally different from still image retargeting and cannot be solved solely by augmenting image-based methods with temporal constraints. The reason for this is twofold: (i) In video, motion and temporal dynamics are the core considerations and must be explicitly addressed; simply smoothing the effect of the per-frame retargeting operator along the time axis, as was done in most previous works, cannot cope with complex motion flow and results in waving and flickering artifacts; (ii) Prominent objects often cover most of the image, in which case any image-based retargeting method reaches its limit, since retargeting is simply impossible without removing or distorting important content. Even if each individual frame does contain some disposable content, the trajectories of the important objects often cover the en-

tire frame space. This makes it impossible to simultaneously preserve the shape of the important objects and retain temporal coherence. All recent methods (e.g., [Wolf et al. 2007; Zhang et al. 2008; Krähenbühl et al. 2009; Wang et al. 2009]) then degenerate to simple linear scaling. Content-aware cropping (such as [Liu and Gleicher 2006; Deselaers et al. 2008]) produces large virtual camera motions and possible loss of salient content.

In this work, we present a video retargeting algorithm that rethinks the problem from the temporal point of view and explicitly puts motion information at the basis. We analyze the motion flow of the entire video sequence and design a simple and concise retargeting framework that achieves temporally coherent and visually-faithful video resizing. Given the fundamental spatial limitation, our first contribution is to give up the notion that all important content must be preserved in *every* frame. Instead, we employ motion information to determine temporal *persistence* of video content. This provides a new cropping criterion, where salient objects may be removed from some frames, but we make sure that they persist in at least a minimal period of time, such that all important content is visible at some point in the video. Our second contribution is a temporal energy for video warping, which is designed to preserve the motion flow and ensures temporally-coherent retargeting. The first step in this direction was done by Wang et al. [2009], as they attempted to consistently resize moving objects and preserve camera motion. However, their method required separation of camera and dynamic motion, which is difficult to impossible for videos with complex dynamic scenes, perspective effects and significant depth variation. We show that, surprisingly, using much simpler temporal constraints it is possible to handle arbitrary motion and parallax effects without any need for camera estimation and alignment or object segmentation.

When the video is crowded with visually important content (as in Figure 2), our temporal persistence based cropping is beneficial over warping, since it introduces no spatial distortion artifacts. On the other hand, when parts of the video do contain unimportant homogeneous regions, content-aware warping is advantageous since it can utilize these regions to hide distortion, even if the regions reside in the middle of the frame (whereas cropping can only remove areas from the periphery). Our third contribution is therefore to optimally combine cropping and warping in one framework, where the algorithm automatically balances between the two depending on the video content (see Figure 1). We achieve this by defining a *critical region* which must not be cropped in each frame. We then apply content-aware warping to the video using our new temporal coherence energy, while constraining the critical regions to remain within the target video cube. Effectively, where previous methods constrained the boundaries of each frame to touch the boundaries of the target cube, we instead only ask to place the critical regions inside the cube, without a priori constraining the frame size. The variational optimization of the warping function then automatically decides how the video contents is transformed and how much of it is left outside of the target cube to be cropped out.

We demonstrate the effectiveness of our retargeting method on numerous challenging examples, and compare our technique to the state-of-the art resizing systems. To evaluate the performance of our framework, we conducted a user study with 96 participants, which showed strong preference of our method over recently published techniques.

2 Related work

Image retargeting. Content-aware image retargeting methods can be roughly classified into discrete and continuous techniques [Shamir and Sorkine 2009]. Discrete methods regard images as collections of pixels and decide which pixels to remove (or duplicate) in order to obtain the target aspect ratio. Cropping techniques [Chen et al. 2003; Liu et al. 2003; Suh et al. 2003; Santella et al. 2006] cut rectangular regions adjacent to the image boundary while trying to avoid removing salient objects. Seam carving [Avidan and Shamir 2007; Rubinstein et al. 2008] removes or duplicates contiguous but not necessarily straight chains of pixels that pass through homogeneous regions of the image. Multi-operator frameworks [Rubinstein et al. 2009; Dong et al. 2009] interleave seam carving, linear scaling and cropping operators while optimizing an image similarity measure. While this strategy combines the advantages of the individual operators, it is very costly because of the evaluation of a sophisticated similarity measure and exponential dependence on the number of operators used. Some recent works employ patches instead of individual pixels and preserve patch coherence between the source and target images, enabling automatic removal of repetitive patterns [Cho et al. 2008; Simakov et al. 2008; Barnes et al. 2009]. The ShiftMap method [Pritch et al. 2009] removes entire objects at a time, alleviating the discontinuity artifacts of pixel-based carving.

©ARS Film Production

Figure 2: *Left: original frames. Right: retargeted frames. Removing some content is necessary when there are too many important objects within the video. With our method, each prominent object remains visible at least for a period of time when retargeting. Top: the people on the far right will remain visible for a while and can thus be cropped in this frame. Combining with warping allows to retain more salient information and win some space by distorting background content.*

Continuous retargeting methods [Gal et al. 2006; Wolf et al. 2007; Wang et al. 2008; Zhang et al. 2008; Karni et al. 2009; Krähenbühl et al. 2009; Zhang et al. 2009] use variational formulation to design warping functions, such that the shape of salient regions is preserved while homogeneous regions are squeezed or stretched. The warps can be discretized at pixel level or coarser, to trade quality for efficiency. Continuous approaches tend to produce smoother results than discrete ones and have more flexibility w.r.t. the optimized objective; however, in some situations removing entire image parts proves better than distorting them. All image retargeting methods have the fundamental spatial limitation: if there is not enough unimportant content in the image, salient objects must be distorted or removed.

Video retargeting. As previously mentioned, most video retargeting works proceed by extending per-frame image-based techniques with some temporal considerations. Cropping methods [Fan et al. 2003; Liu and Gleicher 2006; Deselaers et al. 2008] produce controlled virtual camera motions (such as pan or zoom) and/or artificial scene cuts. Depending on temporal dynamics of the video, the introduced virtual camera motion may be quite large; additionally, important objects might be discarded completely, or at least for a long period of time.

Figure 3: *We define critical regions using vertical lines when narrowing videos. The content within the regions is constrained to persist after resizing. Notice that the sizes of critical regions may be different among frames and not all the regions outside will be necessarily discarded.*

Image resizing methods were extended to video by constraining temporally-adjacent pixels to transform similarly [Wolf et al. 2007; Rubinstein et al. 2008; Zhang et al. 2008; Krähenbühl et al. 2009]. Continuous methods formulate this as an energy term which penalizes strong partial derivatives of the warp w.r.t. time. The seam carving of [Rubinstein et al. 2008] is extended to video by constrained surface carving using graph-cuts. Due to camera and dynamic motion, temporally-adjacent pixels do not necessarily contain corresponding objects, so that objects may deform inconsistently between frames, resulting in waving artifacts. Wang et al. [2009] addressed this temporal coherence problem by explicit detection of camera and object motions; however, their frame alignment assumes each frame to be a plane and cannot handle parallax. The spatial limitation affects all current video resizing methods: if salient objects cover the entire frame space, their temporally-consistent resizing degenerates into linear scaling. By combining warping with temporally-based cropping, we utilize degrees of freedom in the time dimension to overcome this spatial limitation.

3 Motion based video retargeting

Our video retargeting algorithm works by utilizing the knowledge about pixel motions. The motion information implies inter-frame pixel correspondence, which allows our system to (partially) crop out temporally persistent content, as well as to consistently warp corresponding objects throughout the video. In the following we describe both the crop and the warp components of the algorithm and show how to optimally combine them together to minimize visual artifacts in the retargeted media. Since our approach heavily relies on accurate motion information, we employ the state-of-art optical flow method of [Werlberger et al. 2009] to determine the movement of each pixel between neighboring frames. We denote the flow of pixel i by $\mathbf{f}_i$.

3.1 Persistence-based video cropping

Typical video content persists over a sequence of several frames. This provides additional freedom to the retargeting operation and leads us to a novel cropping criterion: even the content which is visually salient in any individual frame can be cropped from some frames, as long as it remains visible for a sufficient amount of time.

Critical regions. We determine the regions which may be cropped by looking at the inverse problem, i.e., by defining a *critical region* in each video frame whose content may not be removed. Critical content is invisible in neighboring frames, or consists of active foreground objects that exhibit significant motion. We use optical flow to define these criteria and compute the critical region of the entire video cube; content outside the critical regions can be potentially discarded. Specifically, when narrowing a video, we look for critical columns of pixels which should not be removed, and when widening we look at critical rows; for brevity we only mention narrowing from now on. We have two criteria for a critical column: (i) its content has just entered the frame or is about to disappear in the next frames, so it is not temporally persistent; (ii) the column contains actively moving foreground objects. We then define critical regions as the areas between the leftmost and the rightmost critical columns.

The horizontal component of the optical flow indicates whether content moves in or out in the next frame; we thus take the average flow vector of each pixel column and test whether it came from any of the previous k_1 frames and will remain visible in the next k_2 frames ($k_1 = k_2 = 30$ in our experiments). If these conditions do not hold, the column is marked as critical. Columns that contain actively moving foreground objects (objects that move independently of the camera motion) are determined by the entropy of the column's flow. To compute the entropy, we quantize the flow vectors $\mathbf{f}_i$ ($i \in C$ where C is the given column pixels) using the common fan chart scheme, where longer vectors are quantized into more levels (tiny flow vectors typically come from noise and do not require as many quantization bits). Let $I(\mathbf{f}_i)$ denote the integer value associated with the flow vector $\mathbf{f}_i$ after quantization:

$$I(\mathbf{f}_i) = 2^k + \Big\lfloor \frac{\theta(\mathbf{f}_i)}{2\pi/2^k} \Big\rfloor, \quad \text{with} \quad k = \lfloor 0.5\mathcal{L}(\mathbf{f}_i) \rfloor, \tag{1}$$

where $\mathcal{L}(\mathbf{f}_i)$ and $\theta(\mathbf{f}_i)$ denote the length and orientation of $\mathbf{f}_i$, respectively. The rationale of this formula is as follows: a fan chart is comprised of concentric rings of equal width, where the outer radius of ring k is $2(k+1)$. Ring k is divided into 2^k equal sectors; each sector spans $2\pi/2^k$ radians. All sectors are consecutively numbered starting with the innermost one. Imagine placing the origin end of the flow vector at the origin of the chart; Eq. (1) then computes the sector index in which the tip of the vector will be. Specifically, $k = \lfloor 0.5\mathcal{L}(\mathbf{f}_i) \rfloor$ is the corresponding ring number and $\lfloor \theta(\mathbf{f}_i)/(2\pi/2^k) \rfloor$ is the particular sector we land in on the ring.

Given the quantized flow values, we then compute the histogram of $I(\mathbf{f}_i)$'s and define flow probabilities $P(\mathbf{f}_i)$ (the heights of the histogram bins normalized by the total integral of the histogram), such that the entropy of column C is

$$H(C) = -\sum_{i \in C} P(\mathbf{f}_i) \cdot \log_2 P(\mathbf{f}_i). \tag{2}$$

We consider columns with flow entropies larger than $0.7H_{max}$ as critical, where H_{max} is the maximal possible entropy (occurring when the flows are uniformly distributed). Figure 3 shows an example of the boundaries of detected critical regions. Note that the crop boundaries serve as constraints, or cropping guides in our system, and not all contents outside will be necessarily fully cropped; the exact amount of cropping depends on the combination with the warping operation and temporal coherence constraints, as we shall see next. Therefore, explicit extraction of foreground objects is not necessary in our system since the flow entropy is a sufficiently good indicator.

original frame [Krähenbühl et al. 2009] [Wang et al. 2009] our results

Figure 4: *Consistently resizing temporally adjacent pixels in the method of [Krähenbühl et al. 2009] introduces waving artifacts when prominent objects move a lot. The camera estimation of [Wang et al. 2009] fails due to lack of reliable feature correspondences. As a result, all local regions are squeezed equally to preserve temporal coherence, similarly to linear scaling.*

3.2 Temporally coherent video warping

Our video retargeting framework is based on a continuous warping function computed by variation optimization [Shamir and Sorkine 2009], and the cropping operation is incorporated by adding constraints to the optimization. We discretize the video cube domain using regular quad grid meshes and define an objective function in terms of the mesh vertices; minimizing the energy function under certain constraints results in new vertex positions; by interpolating the interior of each quad we then reconstruct the retargeted video. The objective function consists of several terms that are responsible for spatial and temporal preservation of visually important content, as well as temporal coherence; we describe these terms next.

Notation. We represent each video frame t using a grid mesh $\mathbf{M}^t = \{\mathbf{V}^t, \mathbf{E}, \mathbf{Q}\}$, where $\mathbf{V} = \{\mathbf{v}_1^t, \mathbf{v}_1^t, ..., \mathbf{v}_n^t\} \subseteq \mathbb{R}^2$ is the set of vertex positions, $\mathbf{E}$ and $\mathbf{Q}$ denote the edges and quad faces, respectively (the connectivity is the same for all frames). The new deformed vertex positions are denoted by $\mathbf{v}_i^{t'} = (x_i^{t'}, y_i^{t'})$; these are the variables in the optimization process. We sometimes drop the superscript t and simply use $\mathbf{v}_i, \mathbf{v}_i'$ when referring to vertices of a single frame to simplify the notation. We denote the target video size by (r_x, r_y, r_z), where r_x, r_y is the target resolution and r_z is the number of frames (which remains unchanged). Conceptually, our goal is to transform the input video cube into the target cube dimensions (without altering the time dimension).

Cropping constraints. Previous warping methods explicitly prescribed the positions of all corner vertices in each frame to match the target resolution. We, instead, design a warp that makes sure that all critical regions (as defined in Section 3.1) are transformed *inside* the target video cube dimensions (r_x, r_y, r_z). Non-critical regions at the peripheries of the video may be transformed outside of the target cube and will thus be cropped out.

Let $\mathbf{v}_\ell^t$ and $\mathbf{v}_r^t$ denote the mesh vertices closest to the top-left and bottom-right corners of the critical region in frame t, respectively (the vertices are chosen conservatively such that the critical region is contained between them). To force the critical region inside the target cube, we must satisfy

$$\begin{aligned} x_\ell^{t'} \geq 0, \quad x_r^{t'} \leq r_x, \\ y_\ell^{t'} \geq 0, \quad y_r^{t'} \leq r_y, \quad \text{for all} \quad 0 \leq t \leq r_z, \end{aligned} \tag{3}$$

Note that by design, our warping function is temporally coherent (see next section) and we do not need to design separate constraints for the temporal coherence of the cropping region.

Spatial content preservation. To preserve the shape of visually important objects in each frame, we employ the conformal energy as in [Zhang et al. 2009], described next for completeness. Each quad is to undergo a deformation which is as close as possible to similarity. Let $\mathbf{v}_{i_1}, \mathbf{v}_{i_2}, \mathbf{v}_{i_3}, \mathbf{v}_{i_4}$ be the vertices of a quad q; similarity transformations in 2D are parameterized by four numbers $[s, r, u, v]$, and we wish to express the best fitting similarity between q and q':

$$[s, r, u, v]_{q,q'} = \underset{s,r,u,v}{\operatorname{argmin}} \sum_{j=1}^{4} \left\| \begin{bmatrix} s & -r \\ r & s \end{bmatrix} \mathbf{v}_{i_j} + \begin{bmatrix} u \\ v \end{bmatrix} - \mathbf{v}'_{i_j} \right\|^2 \tag{4}$$

Since this is a linear least-squares problem, we can write $[s, r, u, v]_{q,q'}^T = (A_q^T A_q)^{-1} A_q^T \mathbf{b}_{q'}$, where

$$A_q = \begin{bmatrix} x_{i_1} & -y_{i_1} & 1 & 0 \\ y_{i_1} & x_{i_1} & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{i_4} & -y_{i_4} & 1 & 0 \\ y_{i_4} & x_{i_4} & 0 & 1 \end{bmatrix}, \quad \mathbf{b}_{q'} = \begin{bmatrix} x'_{i_1} \\ y'_{i_1} \\ \vdots \\ x'_{i_4} \\ y'_{i_4} \end{bmatrix}. \tag{5}$$

Note that the matrix A_q depends solely on the initial grid mesh, and the unknowns are gathered in $\mathbf{b}_{q'}$. By plugging in the expression for $[s, r, u, v]_{q,q'}$ into Eq. (4) we obtain the conformal energy term for quad q as $D_c(q, q') = (A_q (A_q^T A_q)^{-1} A_q^T - I)\mathbf{b}_{q'}$ (see [Zhang et al. 2009] for the derivation details). The total conformal energy term for the entire video is then

$$D_c = \sum_t \sum_{q^t} w_q^t q^t \, D_C(q^t, q^{t'}), \tag{6}$$

where w_q^t is the visual importance of quad q in frame t. The per-frame spatial importance map is obtained from the combination of intensity gradient magnitudes, Itti's saliency measure [Itti et al. 1998] and the robust face detection of [Viola and Jones 2004], similarly to previous warping methods [Wang et al. 2009; Krähenbühl et al. 2009]. The map is normalized to $[0.1, 1.0]$ to prevent excessive shrinkage of unimportant regions.

We also adopt the following energy terms from [Wolf et al. 2007] to prevent strong bending of the mesh grid lines (this is desirable as salient objects tend to occupy connected quads):

$$D_\ell = \sum_t \Big(\sum_{\{i,j\}\in\mathbf{E}_v} (x_i^{t\prime} - x_j^{t\prime})^2 + \sum_{\{i,j\}\in\mathbf{E}_h} (y_i^{t\prime} - y_j^{t\prime})^2 \Big). \quad (7)$$

where $\mathbf{E}_v$ and $\mathbf{E}_h$ are the sets of vertical and horizontal mesh edges.

Temporal coherence preservation. To achieve temporally-coherent video resizing, we design an energy term to preserve the motion information, such that flickering and waving artifacts can be minimized. Given the optical flow we can determine the evolution of every quad q_i^t in the following frame, denoted as p_i^{t+1}. We find the best fitting linear transformation T_i^t such that $T_i^t(q_i^t) \approx p_i^{t+1}$ (we do not include translation in T_i^t since we are only interested in the transformation of the *shape* of each quad, not its precise location). Our goal is to preserve this transformation in the retargeted video, so we formulate the following energy term:

$$D_\alpha(q_i^t) = \|T_i^t(q_i^{t\prime}) - p_i^{t+1\prime}\|^2 \quad (8)$$

Note that the simple energy above encompasses both motions due to camera and independent object motions, without any need to separately handle the two. What remains is to properly formulate it in terms of our unknowns, i.e., the mesh vertex positions. Denote the vertices of p_i^{t+1} by $\mathbf{u}_j^{t+1}$; we represent each of these vertices as a linear combination of the grid mesh vertices $\mathbf{v}_d^{t+1}$ in the immediate vicinity (see Figure 5):

$$\mathbf{u}_j^{t+1} = \sum_d \omega_d \mathbf{v}_d^{t+1}, \quad (9)$$

where ω_d are the barycentric coordinates w.r.t. the quad vertices $\mathbf{v}_d^{t+1}$. Now we can properly reformulate (8) in terms of the $\mathbf{v}_i{}'$s:

$$D_\alpha(q_i^t) = \sum_{(j,k)\in\mathbf{E}(q_i^t)} \|T_i^t(\mathbf{v}_j^{t\prime} - \mathbf{v}_k^{t\prime}) - (\mathbf{u}_j^{t+1\prime} - \mathbf{u}_k^{t+1\prime})\|^2, \quad (10)$$

where $\mathbf{E}(q_i^t)$ is the set of edges of quad q_i^t.

Note that there may be a set of quads $\mathbf{Q}_\beta^t$ which the flow takes outside of the video frame. For such quads, we simply constrain their temporally adjacent quads to be similar after resizing, using the following term:

$$D_\beta(q_i^t) = \sum_{(j,k)\in\mathbf{E}(q_i^t)} \|(\mathbf{v}_j^{t\prime} - \mathbf{v}_k^{t\prime}) - (\mathbf{v}_j^{t+1\prime} - \mathbf{v}_k^{t+1\prime})\|^2. \quad (11)$$

Let $\mathbf{Q}_\alpha^t = \mathbf{Q}^t \setminus \mathbf{Q}_\beta^t$. The overall temporal coherency energy is

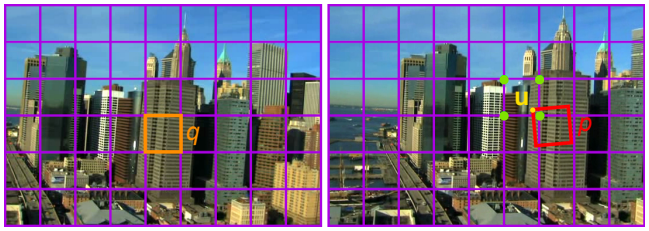
$$D_t = \sum_t \sum_{q_i^t\in\mathbf{Q}_\alpha^t} D_\alpha(q_i^t) + \sum_t \sum_{q_i^t\in\mathbf{Q}_\beta^t} D_\beta(q_i^t). \quad (12)$$

The above energy preserves temporal coherence of corresponding objects using local constraints, which means that inconsistency could accumulate among frames. To handle this problem, it is possible to preserve corresponding quads among farther frames to slow down the error accumulation. Specifically, in Eq. (8), we look at q_i^t and its corresponding quad $p_i^{t+\lambda}$ instead of q_i^t and p_i^{t+1} if their motions are similar ($\lambda = 5$ in our implementation). We noticed however, that allowing slightly inconsistent resizing is reasonable because small changes in objects' shapes are inconspicuous, especially when the camera or objects are moving.

Note that so far the energies were only concerned with the shape of the resized quads, while globally the video frames were allowed to slide, effectively creating an additional "virtual" camera motion. Although such motion may be unavoidable, it is desirable to minimize it since artists usually use camera movement to convey the story, and thus it should be respected as much as possible. We therefore pick an anchor vertex (we chose the top left vertex $\mathbf{v}_0'$) and constrain its position to change smoothly between neighboring frames. That is, we use the following second-order smoothing term:

$$D_s = n \sum_t \|2\mathbf{v}_0^{t\prime} - (\mathbf{v}_0^{t-1\prime} + \mathbf{v}_0^{t+1\prime})\|^2, \quad (13)$$

where n is the number of mesh vertices (this weight balances the energy term against the other terms that use all mesh vertices and not just a single one).



©Mammoth HD

Figure 5: *The corresponding quads q (shown in orange) and p (red) are determined based on optical flow. In this example, we represent the top left vertex of p, denoted by* $\mathbf{u}$*, using linear combination of the green mesh vertices.*

3.3 Optimized crop-and-warp

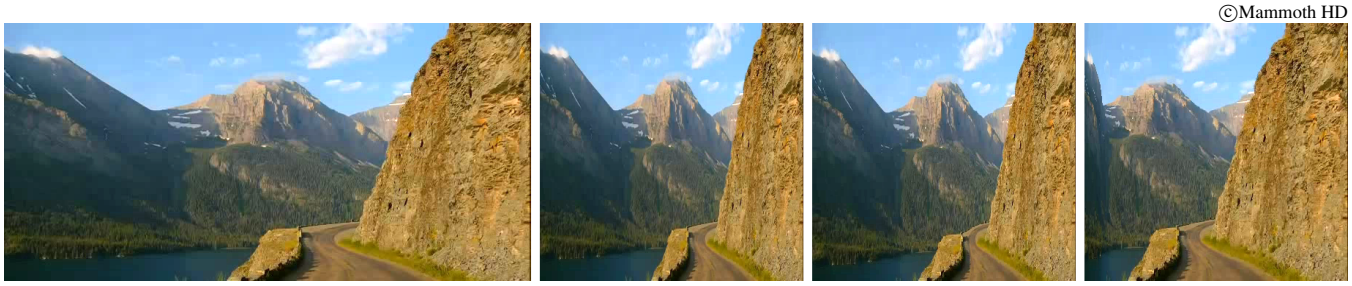
We solve for the deformed grid meshes by minimizing

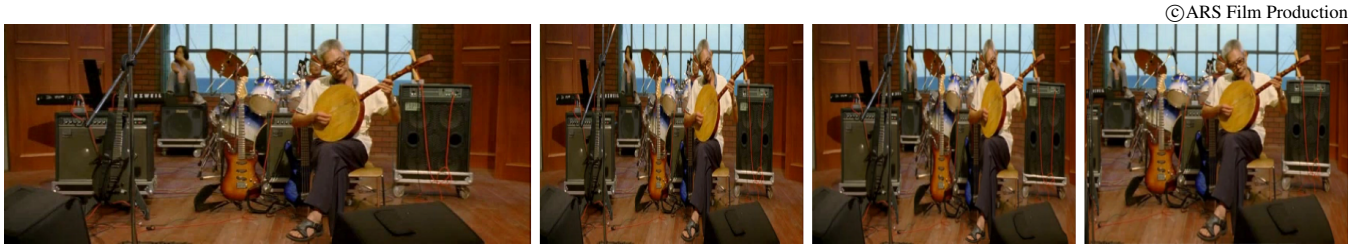
$$D = D_c + D_\ell + \gamma D_t + \delta D_s, \quad (14)$$

where $\gamma = 10$ and $\delta = 1.5$, subject to boundary constraints. The first boundary constraint is the inequality posed by the critical regions (Eq. (3), preventing critical content from being cropped out. We also employ the edge flipping and straight boundary constraints as in [Wang et al. 2008; Wang et al. 2009]: edge flipping is an inequality constraint that prevents self-intersections in the mesh by requiring non-negative length of all mesh edges; straight boundary constraints are linear equations making sure the boundaries of the retargeted frames remain straight (as required for top and bottom boundaries of each frame).

Minimizing the objective function (14) is a linear least-squares problem under some linear constraints and linear inequality constraints, therefore we employ iterative minimization. We start the optimization by placing the leftmost and the rightmost critical columns at the two respective boundaries of the target video cube (note that these columns might reside in different frames; the optimization runs on the entire video cube at once). In each iteration, we solve the linear least-squares problem under the linear equality constraints, which amounts to solving a sparse linear system. We then enforce the detected flipped edges to have zero lengths and also pull the critical columns that turned out to be outside of the target video cube back to the frame boundaries, which effectively results in new equality constraints for the next iteration. Iterations continue until all the inequality constraints are satisfied.

Note that the system matrix changes whenever the constraints change (depending on which inequalities were violated). We apply the GPU-based conjugate gradient solver of [Buatois et al. 2009] with multigrid strategy, which is more memory- and time-efficient than direct solvers in this case. Once the deformed meshes have been computed, we produce the retargeted video by "cutting out" the target cube and interpolating the image content inside each quad (we use linear interpolation, although advanced methods such as EWA splatting [Krähenbühl et al. 2009] could also be employed).





original frame linear scaling [Wang et al. 2009] our results

Figure 6: *The method introduced by Wang et al.[2009] degenerates into linear scaling when the pixel depths vary greatly. Their camera estimation fails to transform corresponding pixels to the same position due to the parallax effect.*

4 Results and discussion

We tested our algorithm on a desktop PC with Duo 2.33 GHz CPU and Nvidia GTX 285 graphics card. We applied the method of [Rasheed and Shah 2003] to cut videos into short clips according to scene changes. Different scenes are retargeted independently, since temporal coherence is not necessary when the contents are disjoint. This strategy improves the performance and memory consumption since the computational complexity is quadratic in the number of unknown vertex positions. To trade quality for efficiency, we typically use grid meshes with 20×20 pixels per quad in our experiments (please see further discussion below). Our retargeting system takes 2 to 3 iterations on average, depending on the video content, when solving the constrained optimization. We use a multigrid strategy to satisfy the inequality constraints on coarser levels in order to improve the performance when deforming finer meshes. Our system can achieve 6 frames per second on average when retargeting a 200-frames video with resolution of 688×288, and the performance naturally drops for larger numbers of frames.

We show some results in Figures 1, 4, 6 and 8 to demonstrate the effectiveness of our algorithm. Please refer to our accompanying and supplemental videos (*MARcomp.mp4* and *SVRcomp.mp4*) for all results and comparisons[1], especially as the temporal effects are difficult to visualize and appreciate in still frames. Note that all results were generated automatically using the default parameters of the algorithm. In some rare cases users may want to manually emphasize important objects; this can be done by segmenting the objects using graph-cut in one frame, and automatically propagating the segmentation to subsequent frames via the optical flow; we show one such example in the accompanying video.

Comparisons. We compared our method with linear scaling, and with the motion-aware retargeting (MAR) of [Wang et al. 2009] and the streaming video retargeting (SVR) of [Krähenbühl et al. 2009], since the latter are the state-of-the-art works on content-aware video resizing. Preceding methods [Wolf et al. 2007; Rubinstein et al. 2008; Zhang et al. 2008] did not handle temporal motion coherence in video resizing and therefore inevitably would not compare favorably with motion-aware methods (this assessment was widely supported by a user study in [Krähenbühl et al. 2009]). Interestingly, the image retargeting techniques of [Dong et al. 2009; Rubinstein et al. 2009] combine cropping and other operators to optimize an image similarity metric; however these methods are already extremely costly for still images and have not been extended to videos in a temporally-coherent manner.

quad	40×40	30×30	20×20	10×10	5×5
time	27 sec.	41 sec.	67 sec.	219 sec.	575 sec.
memory	58 MB	95 MB	225 MB	880 MB	3.5 GB

Figure 7: *Top: resizing results using grid meshes with quads of* 20×20 *(left) and* 5×5 *(right) pixels. Bottom: the timing and memory consumption statistics for various mesh resolutions. This example has* 688×288 *resolution and 208 frames. The experiment was done using a CPU conjugate gradient solver due to the huge memory consumption. Notice that the finest mesh achieves a slightly better result but requires a much heavier computation cost.*

Our main reference for comparison is MAR [Wang et al. 2009] because it explicitly deals with temporal coherence. Since it requires camera alignment which relies on SIFT features, it fails on videos with homogeneous backgrounds (Figure 4). Moreover, when true perspective effects such as parallax are present, their method cannot coherently transform corresponding objects with different depths, in which case the result degenerates to linear scaling (Figure 6). In contrast, our method seamlessly handles all types of motion without requiring camera alignment and therefore succeeds on scenes with arbitrary depth variability and camera motion. Please refer to the accompanying video and *MARcomp.mp4*.

[1] We refer to a set of supplemental results on the project web site at http://graphics.csie.ncku.edu.tw/VideoRetargeting/

Figure 8: *Previous content-aware video resizing methods would fail in this example because the woman overlaps with most of the backgrounds as the camera orbits around her. The temporal constraints of [Wang et al. 2009] cause the method to degenerate into linear scaling. We let parts that are visible long enough to be discarded for some period, which allows preserving the woman's face.*

The pixel-level SVR method of [Krähenbühl et al. 2009] achieves video resizing in real time. To obtain such performance, SVR solves the warping optimization problem on each frame separately, merely constraining temporally-adjacent pixels to be transformed consistently. Temporal coherence is handled by averaging the per-frame spatial importance maps over a window of 5 frames and augmenting them with motion saliency (extracted from optical flow), such that visually prominent and moving objects get higher importance. Yet, per-frame resizing cannot avoid waving artifacts when large camera and dynamic motions are present. In our system, to preserve temporal coherence, we sacrifice real-time efficiency and per-pixel quality and use coarser grid meshes, which allows us to optimize all (or at least more) video frames simultaneously. Please refer to Figure 4 and the supplemental video (*SVRcomp.mp4*) for more comparisons.

Apart from previous state-of-the-art warp-based retargeting methods, we also compared to a manually-generated cropping result in our accompanying video (which should be at least as good as an automatic result). We believe the advantage of warping to be obvious, especially in the challenging examples where the aspect ratio of the video is significantly altered. In our results the width was reduced by 50%; any kind of cropping will suffer from significant object removal or cutting artifacts in this scenario.

It is worth noting that employing finer, or pixel-level mesh resolutions in our method would achieve better results because the saliency and motion information would be more accurately considered. However, the quality improvement when using finer grids is limited, since the contents of each quad is often homogeneous. We have experimented with varying grid resolutions and show some results in Figure 7 and the supplemental video (*multiRes.mp4*). Although the computation and memory costs dramatically increase, the retargeted videos look similar when the meshes are dense enough. We found our choice of 20×20-pixel quads to be a good compromise between quality and performance.

User study. We evaluated our method by conducting a user study with 96 participants coming from diverse backgrounds and ages. We closely followed the study setup of [Krähenbühl et al. 2009], taking the paired comparisons approach [David 1963]: participants were presented with an original video sequence and two retargeted versions side by side, and they were asked to answer which retargeted version they prefer. The users were kept naive about the purpose of the experiment and were not provided with any special technical instructions. We used 6 different videos in the experiment and retargeted each video to 50% width using fully automatic versions of SVR [Krähenbühl et al. 2009], MAR [Wang et al. 2009] and our method. Therefore, for each video there were 3 pairwise comparisons, and each participant was asked to make $3 \times 6 = 18$ comparisons. The videos were selected to include diverse scenes types and motion types: live action footage and CG films, close-ups and wide angle views, single foreground object and several objects, fast moving and slow moving camera, clips with and without parallax effects. We used videos from five commercial feature films and one CG animated short. In our selection, we tried to have as much variety as possible while keeping the number of clips low, since each clip added 3 more comparisons and we could not expect each user to spend more than 20-30 minutes on the experiment. The questions were presented in random order to avoid bias. We obtained a total of $18 \times 96 = 1728$ answers, and each method was compared $2 \times 6 \times 96 = 1152$ times.

↓ was preferred over →	Ours	MAR	SVR	Total
Our method	-	488	508	996
MAR [Wang et al. 2009]	88	-	309	397
SVR [Krähenbühl et al. 2009]	68	267	-	335

Table 1: *Pairwise comparison results of 96 user study participants. A total of 1728 comparisons were performed. Entry a_{ij} in the middle table means: method i was preferred a_{ij} times over method j.*

Table 1 shows the summary of the obtained results, supporting significant preference of our method. Overall, it was preferred in 86.5% $(996/1152)$ of the times it was compared. It was favored over SVR in 88.2% and over MAR in 84.7% of the comparisons. In contrast, SVR was favored only in 29.1% $(335/1152)$ and MAR in 34.5% $(397/1152)$ of the comparisons. The participants tended to agree in their choices: we measured Kendall's coefficient of agreement: $u = 0.356$, which was statistically significant to $p < 0.01$.

Kendall's coefficient of consistence (an indicator of the number of circular triads $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ which means statistical inconsistency of preferences of an individual user) was $\zeta = 1$ for 78% of the users, i.e., they were perfectly consistent. The average consistency coefficient was high, $\bar{\zeta} = 0.94$ with standard deviation of 0.1, and only 3 users had consistency score $\zeta = 0.5$.

A thorough psychophysical testing of video retargeting is out of the scope of this work; we decided to focus on a concise setting, comparing to the two most recent techniques. Since in the user study of [Krähenbühl et al. 2009], the SVR method was shown to be clearly superior to linear scaling and the methods of [Wolf et al. 2007] and [Rubinstein et al. 2008], we did not repeat those comparisons. It would be interesting to conduct a further perceptual study and compare additional retargeting operators, on more video sequences (this would require a more complex experiment design and more participants). It would be also useful to compare with a no-reference study (where the participants do not see the original-size video). We included the reference video in order to check whether users would be bothered by the cropping component of our system, namely the disappearance of important objects for a period of time. However, judging by a small preliminary test we performed, the presence or absence of the original video does not seem to alter the results, because people tend to ignore the reference video and concentrate on the two side-by-side results.

Spatial and temporal distortion propagation. As previously discussed, preservation of temporal behavior and spatial form of salient objects are two conflicting goals. If the trajectory of an important object covers most of the frame, i.e., the object overlaps all background regions at some point in time, preserving temporal coherence means consistently resizing both the object and the entire background, and the only warping operator that achieves this is linear scaling. Our method automatically goes for a temporal tradeoff in this case: it crops some areas for a part of the period they are visible. We demonstrate this in Figure 8, where the camera path orbits around the woman, such that almost all foreground and background regions are correlated. Compared to the pure cropping, our preservation of motion in critical regions guarantees that important objects persist in target videos. In addition, the combination with warping reduces the introduced virtual camera motion. In many examples, there are sufficiently many available homogeneous regions that absorb the warping distortion, such that cropping need not be used to the full extent and is not noticeable; the balance between cropping and warping is conveniently automatically decided by the variational optimization.

Limitations and future work. First and foremost, although our method expands the distortion propagation to the temporal dimension, as opposed to just the spatial domain, retargeting videos with many prominent features and active foregrounds may still produce distortions, both spatially and temporally (please see Figure 9 and our supplemental videos). In such extreme cases it would be necessary to take artistic control over the definition of critical regions in key frames, and let users decide which objects can be permanently cropped out. Similarly, our automatic cropping criterion may be ineffective for extreme tilting camera motion, since prominent objects may need to be cropped forever. Our framework is completely flexible and can admit various cropping constraints, so in the future specific criteria for cropping with tilting motion can be designed. Secondly, our method heavily relies on accurate motion information. Unfortunately, even the best detection methods are sometimes confused by noise and lighting, which would cause our method to preserve the motion of irrelevant parts of the content and/or extend their persistence.Additionally, our method applies coarse grid meshes to retarget videos, and each quad of the mesh may contain several layers of objects moving independently. In this case, the quad transformation may be insufficient to represent the interior motions. Fortunately, continuous warping has high error tolerance, such that the resulting local waving artifacts are less noticeable. Working on a pixel-level grid would eliminate this problem altogether. Finally, due to the computational costs, our method is currently limited in the length and resolution of the videos it can process. It is possible to improve the scalability of the system by using a streaming approach with a sliding window, similarly to [Krähenbühl et al. 2009; Wang et al. 2009], though this method potentially suffers from temporal incoherence. We leave this extension as our future work.

5 Conclusion

We introduced a framework that achieves video retargeting by focusing on motion information. Motion plays the major role in video and distinguishes video retargeting from still image resizing. Our observation is that motion completely dictates the temporal dimension of the retargeting problem, and to a large extent defines the visually prominent content in video. We therefore let the optical flow guide the retargeting process, using it both for spatial components (temporally-coherent warping) and for temporal decisions (persistence based cropping).

Since analysis and optimization over the entire video sequence up to scene cuts is essential to the success of our method, the computational cost is higher than that of real-time systems which only utilize per-frame optimization. Yet we believe that our method provides valuable insights into the video retargeting problem and a non-negligible step forward in terms of the quality of the results, making it suitable for offline high-quality video processing.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. We are grateful to Alexander Hornung and Manuel Lang for insightful discussions and for helping us with the comparisons. We are also grateful to Tino Weinkauf for his comments, to Alec Jacobson for narrating the accompanying video, to Joyce Meng for her help with the video materials and licensing, to the members of Computer Graphics Group/Visual System Lab, National Cheng-Kung University, in particular Kun-Chuan Feng, for helping to conduct the user evaluation, and to all the users who participated in the user study. The usage of the video clips is permitted by ARS Film Production, Blender Foundation and MAMMOTH HD. This work was supported in part by the Landmark Program of the NCKU Top University Project (contract B0008) and by an NYU URCF grant.

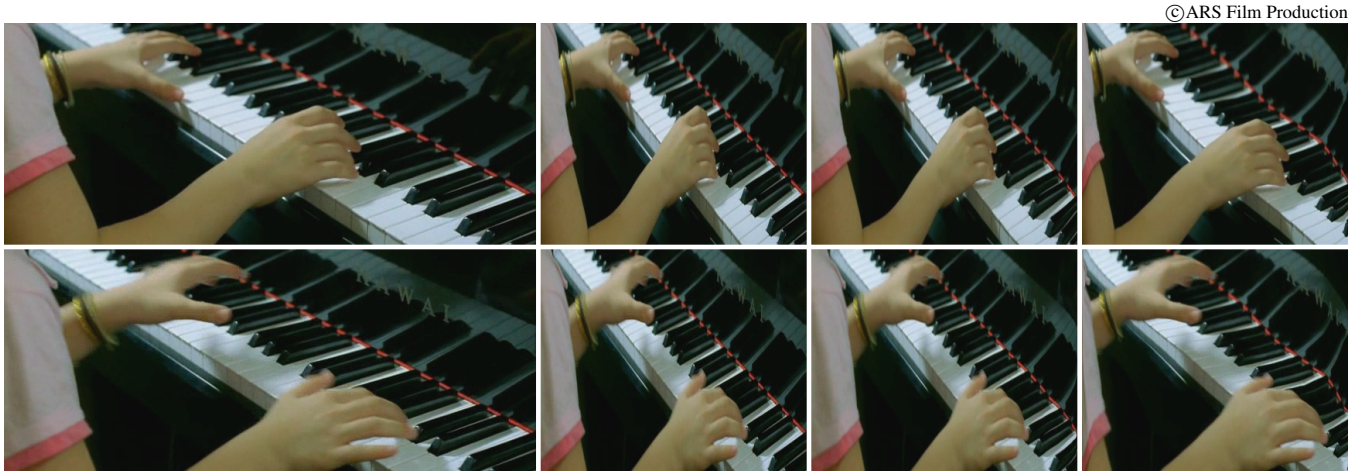
References

AVIDAN, S., AND SHAMIR, A. 2007. Seam carving for content-aware image resizing. *ACM Trans. Graph. 26*, 3, 10.

BARNES, C., SHECHTMAN, E., FINKELSTEIN, A., AND GOLDMAN, D. B. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph. 28*, 3.

BUATOIS, L., CAUMON, G., AND LÉVY, B. 2009. Concurrent number cruncher: a GPU implementation of a general sparse linear solver. *Int. J. Parallel Emerg. Distrib. Syst. 24*, 3, 205–223.

CHEN, L. Q., XIE, X., FAN, X., MA, W. Y., ZHANG, H. J., AND ZHOU, H. Q. 2003. A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal 9*, 4, 353–364.



original frame linear scaling [Wang et al. 2009] our results

Figure 9: *Distortions may occur even though our method strives to hide them in both spatial and temporal domains. In this example, our cropping discards only a small region because objects that are close to video boundaries keep changing among frames and are thus marked as critical. Retargeting the video to 50% of the original width inevitably produces waving and squeezing artifacts. Compared to [Wang et al. 2009], we allow some temporal distortions to better preserve the shape of prominent objects.*

CHO, T. S., BUTMAN, M., AVIDAN, S., AND FREEMAN, W. T. 2008. The patch transform and its applications to image editing. In *CVPR '08*.

DAVID, H. A. 1963. *The Method of Paired Comparisons*. Charles Griffin & Company.

DESELAERS, T., DREUW, P., AND NEY, H. 2008. Pan, zoom, scan: Time-coherent, trained automatic video cropping. In *CVPR*.

DONG, W., ZHOU, N., PAUL, J.-C., AND ZHANG, X. 2009. Optimized image resizing using seam carving and scaling. *ACM Trans. Graph. 28*, 5, 1–10.

FAN, X., XIE, X., ZHOU, H.-Q., AND MA, W.-Y. 2003. Looking into video frames on small displays. In *Multimedia '03*, 247–250.

GAL, R., SORKINE, O., AND COHEN-OR, D. 2006. Feature-aware texturing. In *EGSR '06*, 297–303.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 20*, 11, 1254–1259.

KARNI, Z., FREEDMAN, D., AND GOTSMAN, C. 2009. Energy-based image deformation. *Comput. Graph. Forum 28*, 5, 1257–1268.

KRÄHENBÜHL, P., LANG, M., HORNUNG, A., AND GROSS, M. 2009. A system for retargeting of streaming video. *ACM Trans. Graph. 28*, 5.

LIU, F., AND GLEICHER, M. 2006. Video retargeting: automating pan and scan. In *Multimedia '06*, 241–250.

LIU, H., XIE, X., MA, W.-Y., AND ZHANG, H.-J. 2003. Automatic browsing of large pictures on mobile devices. In *Proceedings of ACM International Conference on Multimedia*, 148–155.

PRITCH, Y., KAV-VENAKI, E., AND PELEG, S. 2009. Shift-map image editing. In *ICCV'09*.

RASHEED, Z., AND SHAH, M. 2003. Scene detection in Hollywood movies and TV shows. In *CVPR '03*, vol. 2, II–343–8.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2008. Improved seam carving for video retargeting. *ACM Trans. Graph. 27*, 3.

RUBINSTEIN, M., SHAMIR, A., AND AVIDAN, S. 2009. Multi-operator media retargeting. *ACM Trans. Graph. 28*, 3, 23.

SANTELLA, A., AGRAWALA, M., DECARLO, D., SALESIN, D., AND COHEN, M. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of CHI*, 771–780.

SHAMIR, A., AND SORKINE, O. 2009. Visual media retargeting. In *ACM SIGGRAPH Asia Courses*.

SIMAKOV, D., CASPI, Y., SHECHTMAN, E., AND IRANI, M. 2008. Summarizing visual data using bidirectional similarity. In *CVPR '08*.

SUH, B., LING, H., BEDERSON, B. B., AND JACOBS, D. W. 2003. Automatic thumbnail cropping and its effectiveness. In *Proceedings of UIST*, 95–104.

VIOLA, P., AND JONES, M. J. 2004. Robust real-time face detection. *Int. J. Comput. Vision 57*, 2, 137–154.

WANG, Y.-S., TAI, C.-L., SORKINE, O., AND LEE, T.-Y. 2008. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph. 27*, 5, 118.

WANG, Y.-S., FU, H., SORKINE, O., LEE, T.-Y., AND SEIDEL, H.-P. 2009. Motion-aware temporal coherence for video resizing. *ACM Trans. Graph. 28*, 5.

WERLBERGER, M., TROBIN, W., POCK, T., WEDEL, A., CREMERS, D., AND BISCHOF, H. 2009. Anisotropic Huber-L1 optical flow. In *Proceedings of the British Machine Vision Conference (BMVC)*.

WOLF, L., GUTTMANN, M., AND COHEN-OR, D. 2007. Non-homogeneous content-driven video-retargeting. In *ICCV '07*.

ZHANG, Y.-F., HU, S.-M., AND MARTIN, R. R. 2008. Shrinkability maps for content-aware video resizing. In *PG '08*.

ZHANG, G.-X., CHENG, M.-M., HU, S.-M., AND MARTIN, R. R. 2009. A shape-preserving approach to image resizing. *Computer Graphics Forum 28*, 7, 1897–1906.