Discontinuous Seam-Carving for Video Retargeting

Matthias Grundmann[1,2] Vivek Kwatra[2] Mei Han[2] Irfan Essa[1]
grundman@cc.gatech.edu kwatra@google.com meihan@google.com irfan@cc.gatech.edu
[1]Georgia Institute of Technology, Atlanta, GA, USA [2]Google Research, Mountain View, CA, USA
http://www.cc.gatech.edu/cpl/projects/videoretargeting

Abstract

We introduce a new algorithm for video retargeting that uses discontinuous seam-carving in both space and time for resizing videos. Our algorithm relies on a novel appearance-based temporal coherence formulation that allows for frame-by-frame processing and results in temporally discontinuous seams, as opposed to geometrically smooth and continuous seams. This formulation optimizes the difference in appearance of the resultant retargeted frame to the optimal temporally coherent one, and allows for carving around fast moving salient regions. Additionally, we generalize the idea of appearance-based coherence to the spatial domain by introducing piece-wise spatial seams. Our spatial coherence measure minimizes the change in gradients during retargeting, which preserves spatial detail better than minimization of color difference alone. We also show that per-frame saliency (gradient-based or feature-based) does not always produce desirable retargeting results and propose a novel automatically computed measure of spatio-temporal saliency. As needed, a user may also augment the saliency by interactive region-brushing. Our retargeting algorithm processes the video sequentially, making it conducive for streaming applications.

1. Introduction

Video retargeting has gained significant importance with the growth of diverse devices (ranging from mobile phones, mobile gaming and video devices, TV receivers, internet video players, *etc*.) that support video playback with varying formats, resolutions, sizes, and aspect ratios. Video retargeting resizes the video to a new target resolution or aspect ratio, while preserving its salient content.

Recent approaches to video retargeting aim to preserve salient content and avoid direct scaling or cropping by removing "unwanted" or redundant pixels and regions [1, 13]. Such a removal (or carving) of redundant regions results in complex non-euclidean transformations or deformations of image content, which can lead to artifacts in both space and time. These artifacts are alleviated by enforcing spatial and temporal consistency of salient content in the target video. In this paper, we propose an algorithm for video retargeting that is motivated by seam carving techniques [1, 13] and augments those approaches with several novel ideas.

Our treatment of video is significantly different than the *surface* carving approach of [13]. We observe that *geometric* smoothness of seams across the video volume - while sufficient - may not be necessary to obtain temporally coherent videos. Instead we optimize for an *appearance*-based temporal coherence measure for seams. We also extend a similar idea to spatial seams, which allows them to vary by several pixels between adjacent rows (for vertical seams). Such a formulation affords greater flexibility than continuous seam removal. In particular, the seams can circumvent large salient regions by making long lateral moves and also jump location over frames if the region is moving across the frame (see Fig. 6a).

To improve the quality of spatial detail over seams as pixels are carved, we propose to use a spatial coherence measure for the visual error that gives greater importance to the *variation* in gradients as opposed to the gradients themselves. This improves upon the forward energy measure of [13]. We demonstrate the effectiveness of this formulation on image resizing applications as well.

Saliency contributes significantly to the outcome of any video retargeting algorithm. Avidan et al. [1] noted that no "single energy function performs well across all images". While we mostly rely on a simple gradient-based saliency in our examples, we also show results that use an alternative fully automatic definition of saliency. This novel definition of saliency is based on the image based approach of [11]. To achieve temporal coherence between frames, we segment the video into spatio-temporal regions and average the frame-based saliency over each spatio-temporal region. We also provide examples generated by user-supplied weighting of spatio-temporal regions. We employ the segmentation algorithm of [5], extended to video volumes [7], for computing spatio-temporal regions, but could have also used segmentations from [8, 12, 16]. In principle, our method is not limited to a single definition of saliency or a specific video segmentation algorithm. While the use of spatio-temporal saliency improves our results considerably we will show that even on per-frame, gradient-based saliency our algorithm outperforms existing approaches.

An additional advantage of our resizing technique is that it processes the video sequentially, *i.e.* on a frame-by-frame basis, and therefore is scalable to arbitrarily long or streaming videos. This allows us to improve the computation time by a factor of at least four compared to the fastest re-

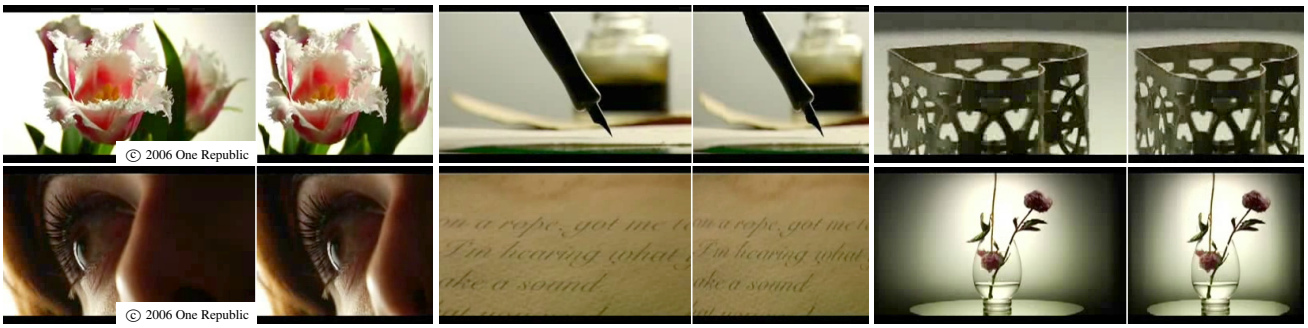


Figure 1: Six frames from the result of our retargeting algorithm applied to a sub-clip of "Apologize", ©2006 One Republic. Original frames on left, retargeted results on right. We use shot boundary detection to separate the individual shots before processing.

ported numbers to-date and achieve performance of about two frames per second.

2. Related Work

The use of seam carving for image resizing was introduced by Avidan and Shamir [1] and later extended for video retargeting by Rubinstein et al. [13]. *Seams* are vertical or horizontal chains of pixels that are successively removed from or added to an image to change its width or height, respectively. To preserve content, seams are chosen as least energy paths through the image. In video, seams are generalized to surfaces that carve through the spatio-temporal volume. Space-time surface carving is also used by Chen and Sen [2] for video summarization. An issue with space-time carving is the memory required for processing video volumes, which is usually addressed by approximation techniques: [2] carve the video in small chunks, while [13] take a banded multi-resolution approach; both use a graph cut algorithm to solve for the surface.

Seam carving is very effective but needs external saliency maps in cases where salient objects lack texture. Wolf et al. [19] present a video retargeting technique that combines automatic saliency detection with non-uniform scaling using global optimization. They compute a saliency map for each frame using image gradients as well as face and motion detection. In contrast, we treat the detection of saliency itself as an orthogonal problem. Primarily, we use per-frame gradient-based saliency similar to [1] but we also generate a temporally coherent saliency based on space-time regions derived from the image-based approach of [11]. We examine the difference of both saliency definitions in Fig. 10 and our video.

Other methods that use optimization for generating visual summaries include [15, 17, 18]. Optimization methods use constraints based on the desired target size. Therefore, they need to be re-run for each desired size. In contrast, seam or surface carving approaches as our proposed algorithm and [1, 13] allow retargeting to the chosen size in real-time. Preventing aliasing artifacts in retargeting was recently addressed by [9] by using a warping technique known as EWA splatting. While producing good results, the approach is mainly constraint to static cameras (*e.g.* line constraints are not tracked).

Gal et al. [6] present a feature-aware texture mapping technique that avoids distorting important features, supplied as user-specified regions, by applying non-uniform warping to the texture image. This is similar to our approach of using regions for saliency. However, our automatic segmentation-aided region selection method scales to video. For video segmentation, we build upon Felzenszwalb and Huttenlocher's graph-based image segmentation [5, 7]. However other video segmentations techniques such as [12] could also be used.

Automatic pan-and-scan and smart cropping have been proposed by [3, 10, 4]. Recently, [14] introduced a method to find an optimal combination of cropping, non-isotropic scaling and seam carving for image retargeting w.r.t. a cost measure similar to [15]. The approach is extended to video by applying the method to key-frames and interpolating the operations between them. We demonstrate equivalent results using our approach and compare to [14].

3. Video Retargeting by Seam Removal

Our video retargeting algorithm resizes a video by sequentially removing seams from it. Seams are 8-connected paths of pixels with the property that each row (vertical seams) or each column (horizontal seams) is incident to exactly one pixel of the seam. Hence removing or duplicating a vertical seam changes the width of a frame by exactly one column. Alternating N times between seam computation and removal for a $w \times h$ frame yields N *disjoint* seams, effectively computing a content-aware resize for $2N$ target sizes $\{(w+N) \times h\}, \ldots, \{(w+1) \times h\}, \{w \times h\}, \ldots, \{(w-N) \times h\}$. This is in contrast to optimization methods that solve for each target size independently. The pre-computed seams enable real-time content-aware resizing as removal

(a) x-t slice (b) Frame 7 (c) Frame 36 (d) Frame 37

Figure 2: Traced x-t slice (at knee height) of a person running from left to right (from Weizmann Action Recognition dataset), obtained using background subtraction. Every vertical surface is a seam in the x-t plane (red) and would intersect with the space-time shape of the person. In contrast our temporally discontinuous solution (green) stays in front of the person (b) and jumps between adjacent frames $(c) \rightarrow (d)$ to overcome spatial distortion.

or duplication of seams only involves fast memory moves.

Rubinstein et al. [13] presented an approach generalizing the seam in an image to a surface in the video volume by extending the image seam carving approach of [1]. The proposed solution for altering the width of the video is a vertical surface. The cross-sections of this surface form a vertical seam in every frame and a *temporal* seam in the $x-t$ plane for any fixed y-location[1]. Therefore, a fundamental property of the surface is that it can only move by at most one pixel-location between adjacent frames.

Consider the case of an object of interest moving from left to right over the video sequence as shown in Fig. 2. Any vertical surface has to start to the right of the object and end to the left of it. In other words, the seam surface is bound to intersect with the object of interest and thereby distort it. This behavior is not limited to this particular case but occurs in general when there is considerable motion in the video perpendicular to the surface – the surface simply cannot keep up with the motion in the video.

In the context of seam carving, temporal coherence is established if adjacent resized frames are aligned like in the original video. If we optimize for temporal coherence *alone*, an obvious solution is to pick the *same* seam for every frame: all pixels that are neighbors along the temporal dimension in the original video will stay neighbors in the resized video. This is akin to non-uniform scaling, where selective columns may be removed (with blending) to shrink the video. However, this by itself will introduce spatial artifacts because in contrast to non-uniform scaling, seams group in non-salient regions instead of being distributed evenly over the columns of a video.

We experimented propagating seams based on tracking non-salient objects in the video. However this does not necessarily lead to good results. In case of vertical seams, if the tracked object does not cover the whole height of the video the propagated seam will intersect with the background at a multitude of different positions resulting in seam that get pulled apart in different directions over time (too fragmented).

Surface carving relaxes the optimal temporal coherence criterion, *i.e.* replicating the same seam in all frames, by allowing the seam to vary smoothly over time. In other words, it imposes a *geometric* smoothness constraint upon the seam solution. While this may be a sufficient condition for achieving temporal coherence, it is not necessary. Instead, we show that, it is sufficient (and less restrictive) to compute a seam in the current frame such that the *appearance* of the resulting resized frame is similar to the appearance obtained by applying the optimal temporally coherent seam. Optimizing against this criterion ensures temporally coherent appearance, but relieves the seams from being geometrically connected to each other across frames, leading to temporally discontinuous seams.

Our algorithm processes frames sequentially as follows. For each pixel in the current frame, we first determine the spatial and temporal coherence costs (S_C and T_C) as well as the saliency (S) cost of removing that pixel. The three cost measures are linearly combined to one measure M, with a weight ratio S_C:T_C:S of 5:1:2 for most sequences. In case of highly dynamic video content we use a ratio of 5:0.2:2. Video clip classification based on optical flow magnitude could automate this choice. We then compute the minimum cost seams w.r.t. M for that frame using dynamic programming, similar to [1]. By removing or duplicating and blending N seams from each frame we can change the width of the video by N columns. Changing the height is achieved by transposing each frame, computing and removing seams, and transposing the resulting frames.

3.1. Measuring Temporal Coherence

Assume we successively compute a seam S^i in every $m \times n$ frame $F^i, i \in 1, \ldots, T$. Our objective is to remove a seam from the current frame so that the resulting $(m-1) \times n$ frame R^i would be visually close to the most temporally coherent one, R^c, where R^c is obtained by reusing the previous seam S^{i-1} and applying it to the current frame F^i.

We use R^c to inform the process of selecting S^i through a look-ahead strategy. For every pixel (x, y), we determine how much the resulting resized frame R^i would differ from R^c if that pixel were removed. We use the sum-of-squared-differences (SSD) of the two involved rows as the measure of temporal coherence, $T_c(x, y)$:

$$T_c = \sum_{k=0}^{x-1} ||F^i_{k,y} - R^c_{k,y}||^2 + \sum_{k=x+1}^{m-1} ||F^i_{k,y} - R^c_{k-1,y}||^2. \quad (1)$$

The temporal coherence cost at a pixel reduces to a per-row difference accumulation that can be determined for every pixel before any seams are computed (see Fig. 3). This allows us to apply the original seam carving algorithm to

[1]Conversely, a horizontal surface forms a horizontal seam in every frame and a temporal seam in the $y-t$ plane for any fixed x-location.

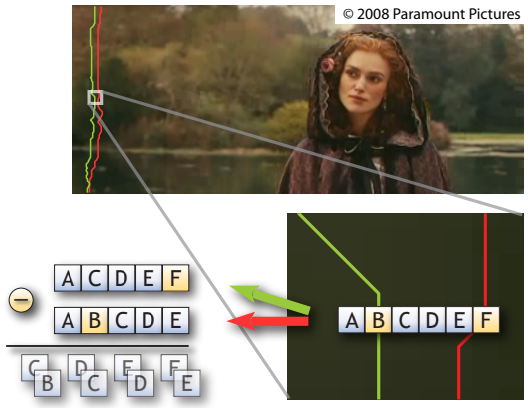


Figure 3: The previous seam S^{i-1} (red) is applied to current Frame F^i. Removing pixel B results in the row $ACDEF$. The optimal temporally coherent seam removes pixel F, so that R^c would contain $ABCDE$. The temporal coherence cost for pixel B is $|C-B|+|D-C|+|E-D|+|F-E|$, which is the SSD between the two rows as well as the sum of gradients from B to F. Original frame from The Duchess, ©2008 Paramount Pictures.

a linear combination of saliency and temporal coherence. It turns out that temporal coherence integrates the gradient along the pixels across which the seam jumps between frames. This is desirable because it means that seams can move more freely in homogeneous regions. Eq.1 can be efficiently computed using two $m \times n$ integral images. The left sum in Eq. 1 will be represented *recursively* by $I^l_{0,y} = 0$, $I^l_{x+1,y} = I^l_{x,y} + ||F^i_{x,y} - R^c_{x,y}||^2$, and the right sum by $I^r_{m-1,y} = 0$, $I^r_{x-1,y} = I^r_{x,y} + ||F^i_{x,y} - R^c_{x-1,y}||^2$, resulting in $T_c = (I^l + I^r)$.

3.2. Measuring Spatial Coherence

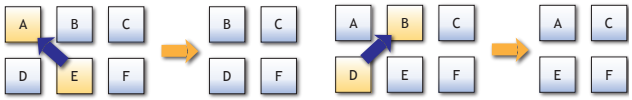
Our look-ahead strategy for measuring temporal coherence may also be applied to the spatial domain. Here, the question is how much *spatial* error will be introduced after removing a seam. The basis of this idea is similar to Rubinstein et al.'s [13] proposed *forward energy*. However, our formulation leads to a more general model, *i.e. piecewise seams*, and is not based on the introduced intensity variation but the *variation in the gradient* of the intensity.

We motivate our spatial coherence measure by examining several different cases in Fig. 4. In (a), there is a step between A and B as represented by the color difference. Removing B yields AC, which exhibits the same step as before, hence no detail is lost[2]. On the other hand, in (b), high frequency detail will be lost on removing B. Removing B in (c) compacts the linear ramp, which is the desired behavior as it compresses the local neighborhood without significantly changing its appearance. In each of these cases, the cost of removing B is well represented by the change in gradient, which is what we use as our measure of spatial coherence, instead of change in intensity.

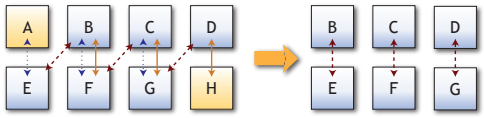
[2]Rubinstein et al.'s forward energy is expressed as a difference in intensity and would be large in this case.

(a) No detail lost (b) Detail lost (c) Linearity preserved

Figure 4: (See in color.) Spatial error if pixel B is removed.



(a) Removing interior pixel (b) Removing border pixel



(c) Piecewise seam cost

Figure 5: Spatial coherence costs: (a) Removing an interior pixel, E w.r.t. A. Bottom row DEF becomes DF, therefore the intensity difference before removing E was $|D-E|+|E-F|$ and is $|D-F|$ afterwards. Between the two rows, the intensity difference was $|A-D|$ and $|B-E|$ and is $|B-D|$ afterwards. (b) Removing a border pixel, here D w.r.t. B. In the bottom row $|D-E|$ becomes $|E-F|$. (c) Summed spatial transition cost for piecewise seams. Consider transition $A \rightarrow H$. We accumulate the change in (LHS) gradient magnitudes before (dotted blue) and after (dashed red) removal (Order: Left to right). We also consider the symmetric case by accumulating the change in RHS gradient magnitudes before (solid orange) and after (dashed red) removal.

Our spatial coherence measure $S_c = S_h + S_v$ consists of two terms, which quantify the error introduced in the horizontal and vertical (including diagonal) directions, respectively, by the removal of a specific pixel. Specifically S_h and S_v are designed to measure the *change in gradients* caused by the removal of the pixel. S_h only depends on the pixel in question and in some sense adds to its saliency, while S_v depends upon the pixel and its potential best seam neighbor in the row above. Therefore S_v defines a spatial transition cost between two pixels in adjacent rows. S_h is defined such that it is zero for the cases (a) and (c) in Fig. 4 and large for case (b). The equations for interior pixels (E in Fig. 5a) and border pixels (D in Fig. 5b) are slightly different, but both measure changes in horizontal gradient magnitude:

$$\text{5a: } S_h(E) = |D-E| + |E-F| - |D-F| \text{ , and}$$
$$\text{5b: } S_h(D) = \big||D-E| - |E-F|\big|.$$

We define S_v to measure the change in vertical gradi-

ent magnitudes when transitioning between a pair of pixels in adjacent rows. We treat the involved pixels in a symmetric manner to avoid giving undue preference to diagonal neighbors. Hence, S_v depends on whether the top neighbor of the pixel in question (say E in Fig. 5a) is its left (A), center (B), or right (C) neighbor. Fig. 5a corresponds to $S_v(E, A)$, where:

$$\begin{aligned} S_v(E,B) &= 0 \\ S_v(E,A) &= \big||A-D|-|B-D|\big|+\big||B-E|-|B-D|\big| \\ S_v(E,C) &= \big||C-F|-|B-F|\big|+\big||B-E|-|B-F|\big|. \end{aligned}$$

Piecewise Spatial Seams: We have shown that in order to achieve temporal coherence, a temporally smooth solution is not necessary; the appearance based measure T_c is sufficient. A natural generalization of this approach is to apply a similar idea to the spatial domain, which would lead to discontinuous spatial seams. For this purpose, we generalize our spatial coherence cost, particularly the transition cost S_v to an accumulated spatial transition cost that allows a pixel to consider not just its three neighbors in the row above but all pixels in that row. An example is shown in Fig. 5c. For a pixel (x_b, y) in the bottom row, the summed spatial transition cost to pixel $(x_a, y-1)$ in the top row (for the case $x_a < x_b$) is:

$$S'_v(x_b, x_a, y) = \sum_{k=x_a}^{x_b-1} |G^v_{k,y} - G^d_{k,y}| + \sum_{k=x_a+1}^{x_b} |G^v_{k,y} - G^d_{k-1,y}|$$

where $G^v_{k,y} = |F_{k,y} - F_{k,y-1}|$ is the vertical gradient magnitude between pixel (k, y) and its top neighbor, while $G^d_{k,y} = |F_{k,y} - F_{k+1,y-1}|$ is its diagonal gradient magnitude with the top right neighbor. The diagonal terms appear because previously diagonal gradients become vertical gradients after seam removal. For the example in Fig. 5c, the *first* term in the equation above will be $|AE-BE|+|BF-CF| + |CG - DG|$, where AE is shorthand for $|A - E|$. The cost for the case $x_a > x_b$ may be defined similarly, while $S'_v(x, x, y) = 0$. In practice, the optimal neighbor x_a typically lies in a window of ~ 15 pixels around x_b, allowing us to reduce the computational cost from $O(m)$ to $O(1)$. Another effect of limiting the search window is that we implicitly enforce seams with a limited number of piecewise jumps in contrast to set of totally disconnected pixels.

Fig. 6 shows examples of both temporally discontinuous and piecewise spatial seams. Fig. 7 demonstrates the effectiveness of our spatial coherence cost in preserving detail. Fig. 9 shows comparisons with image resizing results of [13] (examples from their web page), which use their forward energy measure. Fig. 8 shows a similar comparison for a video example (also from their paper).

(a) Temporally discontinuous seams (b) Piecewise spatial seams

Figure 6: (a) Camera pans to the right. The new seam (green) jumps to the new redundant content on right and avoids introducing artifacts resulting from having to move smoothly through the whole frame. From Sweeney Todd, ©2007 Paramount Pictures (b) Piecewise seams (here neighborhood of 11 pixels) have the freedom to carve around details and therefore prevent artifacts. From The Dark Knight, ©2008 Warner Bros. Pictures.

(a) with S_c (b) w/o S_c (c) [19] (d) [13]

Figure 7: Effect of spatial coherence measure S_c (a) Our algorithm with S_c (without piecewise seams) (b) Our algorithm without S_c (but with [13]'s forward energy); one plane is clearly distorted (c) Our implementation of [19] (d) [13]'s result. Original frame from Valkyrie, ©2007 MGM.

Figure 8: Video retargeting comparison for gradient based saliency. Shown is a single frame from a highway video (top). Our result (bottom-right) is able to preserve the shape of the cars and poles better than [13]'s result (bottom-left). Even the plate on the truck saying "Yellow" is still readable. See accompanying video for complete result.

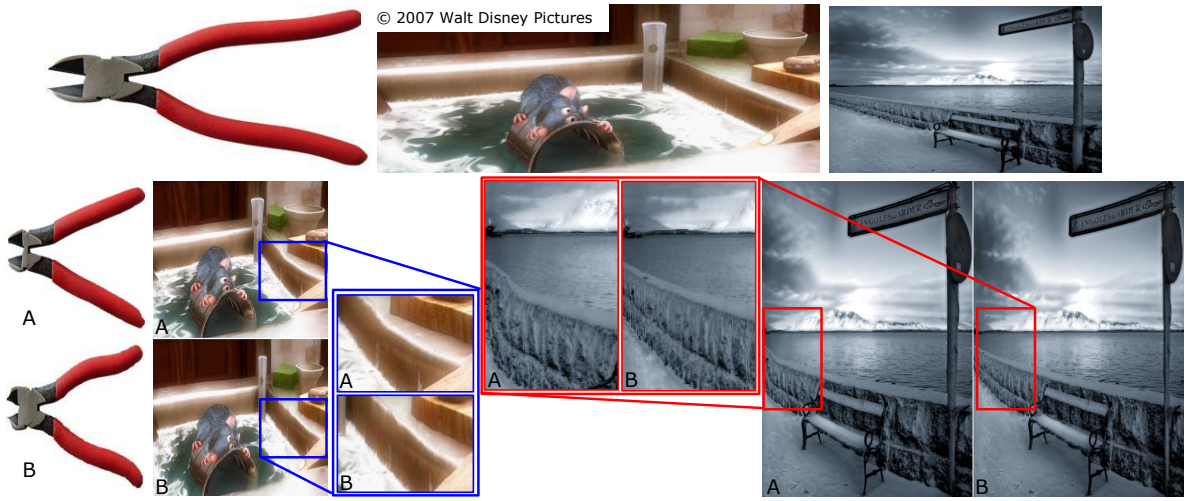


Figure 9: Image retargeting results. Top row shows the original images. In bottom row, images labeled A are [13]'s result, while images labeled B are our results using the novel gradient-variation based spatial coherence cost. In the pliers (left) example, our result better respects the curvature in the handle's shape. For Ratatouille, ©2007 Walt Disney Pictures, (middle) and the snow scene (right), the straight edges are better preserved in our result (shown zoomed in).

4. Automatic spatio-temporal saliency

There are cases where per-frame gradient based saliency[3] is not sufficient. We can employ higher-level techniques such as face detection, motion cues or learned saliency [11], but a major challenge remains in the required temporal coherence for video retargeting. In face detection, for example, the bounding boxes around faces might change considerably between frames or even miss several ones.

We are interested in designing an automatic saliency measure that is temporally coherent as well as aligned with the outlines in the video. The latter requirement is motivated by the fact that local saliency measures do not capture higher-level context and are inherently sensitive to noise. Therefore, we propose to average external per-frame saliency maps over spatio-temporal regions to address both issues. We obtain spatio-temporal regions for video by extending [5]'s graph-based image segmentation to video [7], but any other video segmentation method could be used as well. We build a 3D graph using a 26-neighborhood in space-time with edge weights based on color difference. We then apply the graph-segmentation algorithm to obtain spatio-temporal regions. The effect of applying our method to frame-based saliency maps is shown in Fig.10.

If the underlying frame-based saliency method fails to detect salient content in a majority of frames, the spatio-temporal smoothing fails as well. In this case we offer a user interface that allows highlighting salient and non-salient regions by simple brush strokes, which are then automatically tracked over multiple frames through the spatio-temporal regions (see Fig. 11).

[3]We use the sum of absolute values of the pixel's gradients in our work.

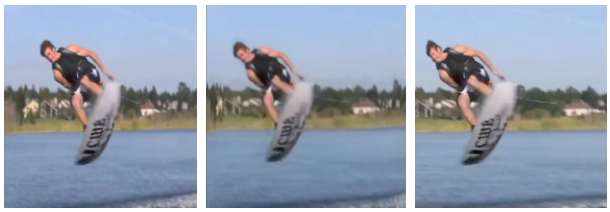
Figure 10: Effect of our spatio-temporal saliency. Left column: Saliency maps computed based on [11] for adjacent frames (top/bottom) independently (white = salient content). Notice the abrupt changes in face, coat and right brick wall. Middle column: Saliency averaged over spatio-temporal regions results in smooth variations across frames. Right column: Effect on video retargeting. Top uses spatio-temporal saliency, bottom uses gradient based saliency. Original frame: 88 minutes, ©2007 TriStar Pictures.

(a) User-selected regions (b) Auto-selected regions

Figure 11: User selects regions in a single frame (a) by roughly brushing over objects of interest (indicated by dashed line). These regions are automatically extrapolated to other frames (b) of the video. See accompanying video. Original frame from 88 minutes, ©2007 TriStar Pictures.

5. Results

We demonstrate our results for video retargeting based on gradient-based saliency and spatio-temporal saliency (auto-



(a) Our result (b) Result from [9] (c) Result from [14]

Figure 12: Comparison to [9] and [14]. Content is highly dynamic (athlete performing 720° turn and fast moving camera). In [9], the background gets squished on the left, the waterfront at the bottom gets distorted, and the result is less sharp overall compared to our result. The approach of [14] distorts the head and essentially crops the frame, while our algorithm compresses the background.

matic as well as user-selected) in the accompanying video. Fig. 12 shows comparisons to other techniques for a highly dynamic video. Fig. 1 and Fig. 13 (top three rows) show frames from example videos that were retargeted using gradient-based saliency. Fig. 13 (bottom row) and Fig. 14 were retargeted by user-selected regions as shown. In both cases it took less than 10 seconds to select the regions.

Our approach provides the user control over determining what regions to carve in case automatic approaches fail. Fig. 14 demonstrates the usefulness of user-selected regions for non-salient content. Fig. 15 shows that we can achieve results comparable to and with sharper detail than [14] – we only used per-frame gradient-based saliency in this case.

6. Conclusion and Limitations

We have presented a novel video retargeting algorithm based on carving discontinuous seams in space and time that exhibits improved visual quality, affords greater flexibility, and is scalable for large videos. We achieve 2 fps on 400x300 video compared to 0.3-0.4 fps for [13] and 0.5 fps for our implementation of [19]. We have presented the novel idea of using spatio-temporal regions for automatic or user-guided saliency. We have also demonstrated the benefits of our novel gradient-variation based spatial coherence measure in preserving detail.

We can handle videos with long shots as well as streaming videos using frame-by-frame processing. However, if spatio-temporal saliency is also employed, then the video length is limited by the underlying video segmentation algorithm, which in our case is $\sim 30 - 40$ seconds.

Fast-paced actions or highly-structured scenes might have little non-salient content. In these cases, just like other approaches, our video retargeting might produce unsatisfactory results as shown in our accompanying video.

The sequential nature of our video retargeting algorithm can occasionally cause the seam in the initial frames to be sub-optimal w.r.t. the later frames. This can sometimes cause several seams to jump their location across time in the same frame, which leads to a visible temporal discontinuity. However, this problem can be alleviated by looking up the saliency information forward in time (around 5 frames) and averaging it with the current saliency.

References

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM SIGGRAPH*, 2007. 1, 2, 3

[2] B. Chen and P. Sen. Video carving. In *Eurographics 2008, Short Papers*, 2008. 2

[3] T. Deselaers, P. Dreuw, and H. Ney. Pan, zoom, scan – time-coherent, trained automatic video cropping. In *IEEE CVPR*, 2008. 2

[4] X. Fan, X. Xie, H.-Q. Zhou, and W.-Y. Ma. Looking into video frames on small displays. In *ACM MULTIMEDIA*, 2003. 2

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2), 2004. 1, 2, 6

[6] R. Gal, O. Sorkine, and D. Cohen-Or. Feature-aware texturing. In *Eurographics*, 2006. 2

[7] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *IEEE CVPR*, 2010. 1, 2, 6

[8] S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. In *IEEE CVPR*, 2001. 1

[9] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross. A system for retargeting of streaming video. In *ACM SIGGRAPH ASIA*, 2009. 2, 7

[10] F. Liu and M. Gleicher. Video retargeting: automating pan and scan. In *ACM MULTIMEDIA*, 2006. 2

[11] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *IEEE CVPR*, 2007. 1, 2, 6

[12] S. Paris. Edge-preserving smoothing and mean-shift segmentation of video streams. In *ECCV*, 2008. 1, 2

[13] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. In *ACM SIGGRAPH*, 2008. 1, 2, 3, 4, 5, 6, 7, 8

[14] M. Rubinstein, A. Shamir, and S. Avidan. Multi-operator media retargeting. In *ACM SIGGRAPH*, volume 28, 2009. 2, 7, 8

[15] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. Summarizing visual data using bidirectional similarity. In *IEEE CVPR*, 2008. 2

[16] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic kernel mean shift. In *ECCV*, 2004. 1

[17] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. Optimized scale-and-stretch for image resizing. *ACM SIGGRAPH ASIA*, 2008. 2

[18] L. Y. Wei, J. Han, K. Zhou, H. Bao, B. Guo, and H. Y. Shum. Inverse texture synthesis. In *ACM SIGGRAPH*, 2008. 2

[19] L. Wolf, M. Guttmann, and D. Cohen-Or. Non-homogeneous content-driven video-retargeting. In *IEEE ICCV*, 2007. 2, 5, 7

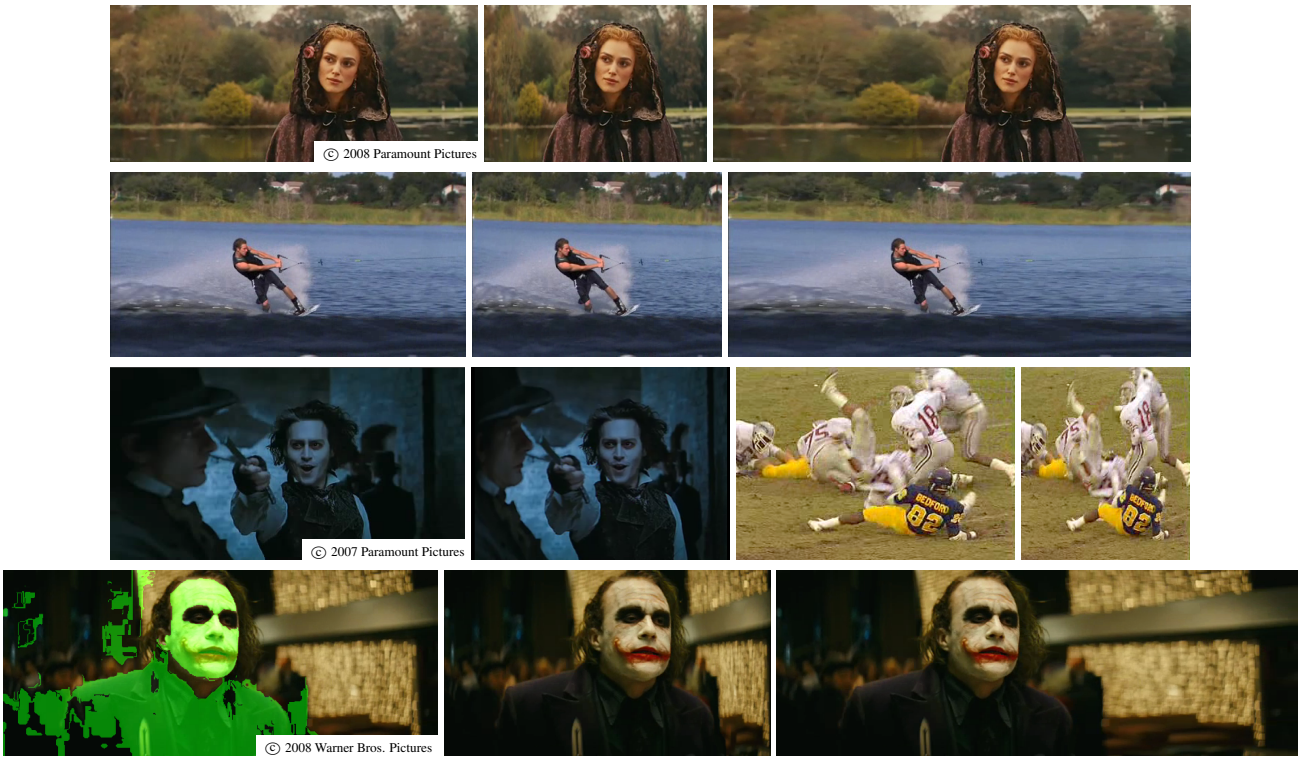
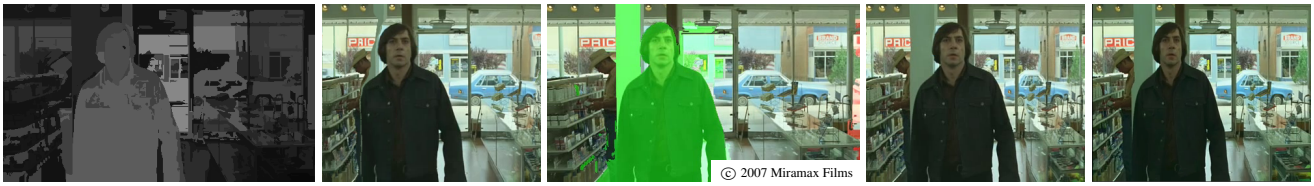


Figure 13: Video retargeting results. Original frame on left. Retargeted result(s) on right. The top three rows show results obtained by our discontinuous seam carving computed on gradient-based saliency. Bottom row shows video retargeted using user-selected regions (marked in green, complexity caused by segmentation errors due to blocking artifacts). Original frames from: The Duchess, ©2008 Paramount Pictures (1st row), Sweeney Todd, ©2007 Paramount Pictures (3rd row), The Dark Knight, ©2008 Warner Bros. Pictures (4th row).



(a) Saliency map (b) Result A (c) User selected regions (d) Result B (e) [13]

Figure 14: Sometimes it is vital to preserve non-salient objects because their removal introduces unpleasant motion. Result A (b) removes the white pillar because it is marked non-salient by the saliency map (a). If we constrain the solution by user-selected regions (c) the pillar is preserved and the outcome is temporally coherent – Result B (d). Please see video for comparison. Compared to [13] (e) our result does not squish the actor and or introduce a bump in the pillar. Original frame from No Country for Old Men, ©2007 Miramax Films.

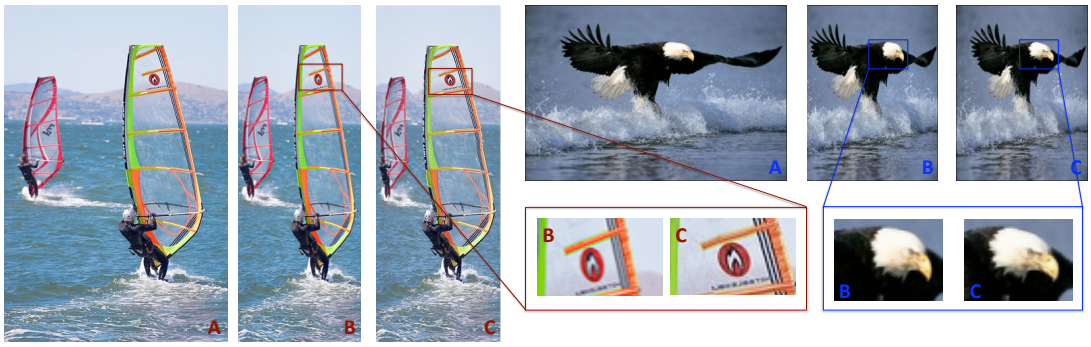


Figure 15: Comparison to [14]. The original image A is resized by the method of [14] using a combination of seam carving, cropping and non-isotropic scaling (B). We achieve similar results (C) using our seam carving *alone* applied to simple gradient-based saliency. Because we avoid scaling and cropping our results have sharper details (see zoomed-in portion).