Gaze-Driven Video Re-Editing

EAKTA JAIN
Carnegie Mellon University, Disney Research Pittsburgh, and University of Florida
YASER SHEIKH
Carnegie Mellon University
ARIEL SHAMIR
Disney Research Pittsburgh and Interdisciplinary Center Israel
and
JESSICA HODGINS
Carnegie Mellon University and Disney Research Pittsburgh

Given the current profusion of devices for viewing media, video content created at one aspect ratio is often viewed on displays with different aspect ratios. Many previous solutions address this problem by retargeting or resizing the video, but a more general solution would re-edit the video for the new display. Our method employs the three primary editing operations: pan, cut, and zoom. We let viewers implicitly reveal what is important in a video by tracking their gaze as they watch the video. We present an algorithm that optimizes the path of a cropping window based on the collected eyetracking data, finds places to cut, and computes the size of the cropping window. We present results on a variety of video clips, including close-up and distant shots, and stationary and moving cameras. We conduct two experiments to evaluate our results. First, we eyetrack viewers on the result videos generated by our algorithm, and second, we perform a subjective assessment of viewer preference. These experiments show that viewer gaze patterns are similar on our result videos and on the original video clips, and that viewers prefer our results to an optimized crop-and-warp algorithm.

Categories and Subject Descriptors: I.2.10 [**Vision and Scene Understanding**]: Video Analysis; I.3.3 [**Picture/Image Generation**]: Viewing Algorithms; I.4.8 [**Scene Analysis**]: Time-varying Imagery

General Terms: Algorithms

Additional Key Words and Phrases: Perceptually-based algorithms, video retargeting, video editing, eyetracking, curve fitting

ACM Reference Format:

Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. 2015. Gaze-driven video re-editing. ACM Trans. Graph. 34, 2, Article 21 (February 2015), 12 pages.
DOI: http://dx.doi.org/10.1145/2699644

Authors' addresses: E. Jain (corresponding author), Y. Sheikh, and J. Hodgins Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213; email: ejain@cise.ufl.edu; A. Shamir, Disney Research Pittsburgh, 4720 Forbes Avenue, Pittsburgh, PA 15213.

DOI: http://dx.doi.org/10.1145/2699644

Fig. 1. We present a gaze-driven algorithm to re-edit widescreen video (e.g., 1.75:1) to smaller aspect ratios via pans, cuts, and zooms. The result computed by our algorithm (in color) is overlaid on the original widescreen frame (in grayscale). Images courtesy The Walt Disney Company.

1. INTRODUCTION

Viewers consume digital content on a wide variety of display devices, ranging from hand-held personal displays such as cellphones or pico-projectors, to large displays such as theater systems. The aspect ratio, size, and resolution of the target display device greatly influence the way filmmakers create and edit a movie. Therefore, when the aspect ratio for display is different from the aspect ratio anticipated at the time of the video's production, a significant modification of the content is required. The challenge is to preserve, as much as possible, the narrative and impact of the original video.

Several methods have been used to modify the size of videos, including uniform scaling, cropping, and letterboxing. Recently, methods for retargeting videos nonuniformly to different aspect

ratios were proposed (see Shamir and Sorkine [2009]). However, these methods are susceptible to noticeable artifacts, such as squeezed shapes and waves, when a large change in aspect ratio is needed. A solution is required that is closer in concept to *re-editing* the movie for the new target display.

Because it is usually impossible to reshoot scenes, change the shooting angle, or add new shots, our challenge is to re-edit the original footage to fit a different display without losing significant content. We employ the three primary editing operations: pans, cuts, and zooms. Panning is used to show different parts of the underlying scene on the screen. A cut allows for a quick shift from one part of the scene to another, when two people are talking, for example. Cropping the widescreen footage gives the effect of "zoom", that is, moving closer to the action. Our method re-edits the given video footage by automatically combining pans, cuts, and zooms. Examples are shown in Figure 1.

This approach of re-editing rather than rescaling has the advantage that it guarantees there will be no distortion in the resulting video. Because the editing operations only remove contiguous portions of the original frame, all scene structure is preserved: lines remain straight and people are not made tall and skinny. At the same time, because this approach makes a hard commitment about what is and is not included in the frame, it depends on reliably determining which parts of the original widescreen frame are integral to the narrative in the original video. We allow the viewers to reveal what is important to the narrative by tracking their gaze.

Artists employ various devices to highlight those regions of the video that are integral to the narrative, including color (the lady wears a striking red dress), motion (the rebel walks opposite to the crowd), semantics (the witness points to the gun), and audio (dialogue or sound effects). These devices attract the viewer's gaze through a combination of bottom-up influences, such as color, motion, or the presence of a human face, and top-down influences such as semantics and narrative.[1] Current saliency algorithms for images or videos primarily capture bottom-up influences and, as a result, often misfire. We expect that research will eventually yield new algorithms that incorporate top-down influences to better predict visual saliency. Rather than waiting for these algorithms, we directly use measurements of the viewers' gaze as input data.

We use recorded gaze data from several human participants as the driving input to a RANSAC algorithm that finds pans, cuts, and zooms while retaining those regions attended to by the participants in the original video. These editing operations mimic the approach of studio professionals when a film shot in widescreen is fit by hand to smaller aspect ratios. The algorithm is able to handle noise in the eyetracking data (a result of individual variation in the gaze patterns across participants, and measurement error). We use RANSAC to search for a path that moves a cropping window through the video cube while maximizing the number of gaze points included. When cuts are required to produce a video with adequate coverage of the eyetracking data, the algorithm simultaneously finds two panning paths and an optimal cut between them. The size of the cropping window is determined from the spread of the gaze samples. The change in size creates the effect of "zoom" in the resulting video.

To evaluate our algorithm, we eyetrack viewers on our results and compare this data to the gaze data captured on the original widescreen videos. For a re-edit to be faithful, viewers must have attended to the same regions in the result as in the original video. This measure allows the evaluation to be performed behaviorally. In our experiments, we find that viewer eye movements are similar before and after the re-edits. We also perform a subjective evaluation of our results by asking users to provide their preferences.

(a) Orignal frame (1.75:1) (b) Scaling (1:1) (c) Cropping (1:1) (d) Letterboxing (1:1)

Fig. 2. Commonly used methods to resize video: (a) Original frame; (b) scaling squeezes the objects and characters; (c) cropping removes content; (d) letterboxing wastes screen space. Still frame from Herbie Rides Again courtesy of The Walt Disney Company.

Contributions. Our main contribution is a method for re-editing video to fit to a nonnative aspect ratio using pans, cuts, and zooms on the original widescreen video. We present a RANSAC-based algorithm that fits a curve to viewers' temporal gaze data. This curve represents the center of the output window, guiding the selection and combination of the re-editing operators (pans and cuts). The size of the window guides the zoom operator. We evaluate our method through a subjective two-alternative forced-choice test of viewer preferences, and a comparison of viewer gaze on pre- and postedited video. Viewer preferences indicate that, amongst re-editing algorithms, our gaze-driven results are preferred to an optimized crop-and-warp method. When viewers are presented with a letterboxed version, they prefer it to a cropped video.

2. RELATED WORK

Resizing a video to fit a nonnative aspect ratio is a controversial task. Cinema enthusiasts often prefer letterboxing, which involves placing a black matte around the original video (Figure 2(d)). Though this method preserves the shape of the original video, it wastes precious screen space, especially on small displays. As a result, there have been several proposed solutions to automatically create nonletterboxed versions of widescreen films. Linear scaling (Figure 2(b)) squeezes or stretches the video to match the size of the new display device, thereby changing the shapes of objects in the video. Center cropping (Figure 2(c)) trims the original video to the desired size, which can result in the "talking noses" artifact.

As a result of these difficulties with simple automatic solutions, skilled editors are called upon to create a "pan and scan" version of a widescreen video. The editor determines the region of the screen that is important to the story for each frame of the video, and moves the cropping window to this region [Wikipedia 2015]. As the "important region" moves across the screen, the cropping window pans across or zooms in and out. The editor could also introduce cuts to avoid panning too fast or too often. Because pan and scan crops away part of the picture, the director relies on the skill and experience of the editor in selecting those regions of the screen that best communicate those narrative and context.

In recent years, researchers have proposed automatic content-aware retargeting methods: resizing the original format video nonuniformly so that it fits the new screen size while minimizing the loss and distortion of visually important content. Two principal classes of methods were introduced for content-aware retargeting [Shamir and Sorkine 2009]. Discrete methods treat the video as a collection of individual pixels, and their goal is to add or remove pixels to achieve the desired size while minimizing an energy measure [Avidan and Shamir 2007; Rubinstein et al. 2008]. Continuous methods treat the video data as a continuous signal that

[1]For a discussion on mechanisms of attention, see Baluch and Itti [2011].

is sampled according to a function, and transform this sampling function to yield an image of the desired size [Wang et al. 2008, 2009; Niu et al. 2010]. Both classes of methods are susceptible to the same types of visual distortions, such as squeezed shapes or broken lines, and temporal incoherence ("waves" in the video).

Thus, more recent approaches have begun to incorporate the classic pan-and-scan operator by allowing for a cropping window [Krähenbühl et al. 2009; Wang et al. 2010, 2011; Xiang and Kankanhalli 2010a, 2010b]. These algorithms rely on computational saliency to identify the regions to discard. Because this identification is not exact, they do not know for sure which regions to leave out. Combining nonlinear scaling with a cropping window allows the operator to discard less, and therefore make fewer wrong decisions, at the expense of distortion and waving artifacts. For example, when two people are talking, saliency algorithms will fire on both faces (because they do not process audio or dialogue) and a combined operator might have to squeeze the two faces (distortion) to fit them inside a smaller aspect ratio because it has no information about which person is more salient.

Liu and Gleicher [2006] automated the pan-and-scan operator for cinematic content based on computed saliency maps. They created a pan by moving a fixed-size window, a zoom by changing the size of a stationary window, and a cut by switching between two stationary windows. Our work generalizes these operators by demonstrating a method to appropriately select combinations of the individual operators to effectively communicate the story. For example, as shown in Figure 1, any single operator would not have captured the entire action.

Deselaers et al. [2008] presented a pan-and-scan-based algorithm that simultaneously searches for the size and position of a cropping window that encloses the relevant regions of the video. They disallow cropping window sizes less than fullscreen, that is, they only allow a zoom out. Perhaps this constraint is meant to avoid cropping out relevant content because they use computational saliency, based on image features such as color and optical flow, that can easily misfire. Their formulation also penalizes nonsmooth trajectories and thus is limited to pans and zooms.

Additionally, there have been several other cropping-window-based approaches [Wang et al. 2004; Tao et al. 2007; El-Alfy et al. 2007; Kopf et al. 2011]. El-Alfy et al. [2007] propose a method similar to ours and that accommodates cuts. These approaches are built for noncinematic videos, such as surveillance videos or sports videos, and they weigh information content more than cinematic guidelines. Our method, on the other hand, is designed for cinematic content, for example, by requiring an ease-in-ease-out profile for the cropping window trajectory.

Our approach to the problem of fitting widescreen video to a smaller aspect ratio is to re-edit the given footage so as to as closely as possible represent the narrative-important regions of the video (as captured by viewer gaze) within the constraints of the desired aspect ratio. It is designed to allow for the simultaneous application of pans, cuts, and zooms. We rely on eyetracking data from viewers to identify which regions of the film are relevant and at what points in time (a teacup might be relevant when the protagonist says that she just drank tea, but might become irrelevant as she describes her plans to assassinate the head of state). With this information, our algorithm can confidently crop away regions that are irrelevant. Thus, this method is inspired by the methodology of human artists, who edit the original widescreen footage keeping the story in mind and make decisions about how to tell this story at a new aspect ratio.

Past work by graphics researchers has used eyetracking data collected when viewers looked at images. Eyetracking has been used for creating painterly renderings [DeCarlo and Santella 2002] and cropping photographs [Santella et al. 2006], and to generate the saliency map input to nonlinear image retargeting methods [Castillo et al. 2011]. In parallel with this work, Katti et al. [2014] utilized eyetracking data on previous frames to predict saliency on subsequent frames for seam-carving-based retargeting of streaming videos. These works essentially used only the spatial locations of viewer gaze, whereas our algorithm utilizes the additional temporal information in gaze data to compute the trajectory of a cropping window.

Eyetracking data, though rich in information, is also very noisy. The noise is a result of measurement error (for example, imperfect calibration) and also a result of variations between individual participants. This second source of noise is reduced for artist-created content because artists actively engage the viewers' attention. Recently, Jain and colleagues [2012] recorded this phenomenon for comic book images. For videos in particular, Dorr et al. [2010] showed that eye movements are significantly more consistent across viewers for Hollywood action movies compared to movies of natural scenes, such as a video taken at a beach, and Mital et al. [2010] validated that motion is a strong attentional cue. Goldstein and colleagues [2007] found that, when human participants watched movie clips, they attended to less than 12% of the screen area more than half the time. This finding suggests that eyetracking data from viewers will reasonably indicate those regions of the video important to the narrative. As eyetracking technologies become less expensive and more easily available (for example, Webcam-based eyetracking [Agustin et al. 2010; Abbot and Aldo 2011; Rudoy et al. 2012]), it will become possible to crowdsource viewer gaze data collection, making our algorithm easier to apply.

3. METHOD

We measure what is important to viewers in a video by recording their gaze as they watch the video. The gaze data is input to a RANSAC algorithm to find a cropping window path through the video cube that encloses maximum saliency while maintaining the specified smoothness characteristics. When saliency shifts sharply from one part of the original widescreen video to another, we cut between two cropping windows. The time location of the cut is optimized within the RANSAC loop, based on the shift in gaze position and the increase in enclosed saliency as a result of the cut. In addition, the spread of the gaze samples in each frame is used to change the cropping window size to create zooms. We now describe the eyetracking procedure and the algorithm to fit pans, cuts, and zooms to this data.

3.1 Data Collection

We selected a variety of clips from three Hollywood films shot in different native formats (2.35:1, 1.82:1, 1.75:1) as the original widescreen videos that would be processed by our algorithm. A categorization of the clips is presented in Figure 6. Six naive participants (1 male, 5 female, ages ranging from 21 to 44 years) were recruited from the university community. All participants had normal vision or wore contact lenses or glasses to correct to normal vision, and were compensated monetarily for their time.

Participants were asked to watch a set of video clips that were resized to be as large as possible on a 19-inch screen (1680×1050), with black letterboxing to preserve the original aspect ratio. The participants sat approximately 18–24 inches from the screen. A visual angle of 1° is approximately 27 pixels at these settings. Before beginning the data collection, participants were asked to adjust the chair to a height and distance comfortable to them; then the system was calibrated. After calibration, they were able to

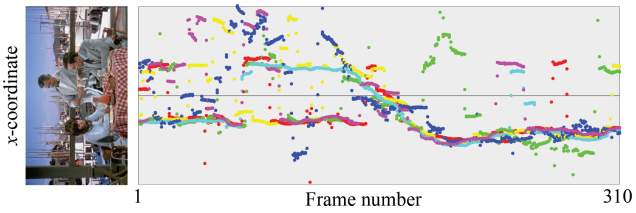

Fig. 3. Viewer gaze is plotted, and an example frame is shown from the corresponding video (each viewer is a unique color). The x-positions are marked on the vertical axis, and frame numbers are on the horizontal axis. Still frame from Herbie Rides Again courtesy The Walt Disney Company.

move their head freely while their eyes were tracked. This setup allowed for a natural viewing situation as no chin rest was required. The stimuli video clips were presented in randomized order. Before each clip, one line of text was displayed to provide context about the story. The eye movements of the participants were recorded with SensoMotoric Instruments' RED eyetracker, running the iViewX software at 60 Hz. Raw data is illustrated in Figure 3.

3.2 Cropping Window

The pan is created through a cropping window that moves across the original widescreen video. It is parametrized by the position of its center (x_i, y_i), $i = 1, 2, \ldots, F$, where F is the number of frames in the video, and its size $\mathbf{D}_i \in \mathbb{R}^2$, for example, $\mathbf{D}_i = [1050, 1680]$. When there is no zoom, the size $\mathbf{D}_i$ is fixed to be the largest window with the specified aspect ratio that fits in the widescreen video. The cropping window path is obtained by computing the x-coordinate of the center of the cropping window, that is, x_i, from the input gaze data.

Naively smoothed or filtered eyetracking data cannot be used to drive the camera's motion because eye movements and camera movements are different in several ways. A panning camera is often employed to reframe a subject so that it remains in the frame while moving. The pan is thought to mimic the smooth motion of the eye as it follows a subject, while keeping the movement minimally noticeable [Katz 1991]. The purpose of eye movements is to process information, which can occur by quickly shifting visual attention (saccades), resting at the same location (fixations), or smoothly following a moving target (smooth pursuit). Thus, we represent the path of the cropping window with B-spline curves that are then fitted to the gaze samples. This approach allows us to enforce the smooth ease-in-ease-out motion that helps keep the pan minimally noticeable. We use a RANSAC algorithm to be robust to noise, including individual variation in eye movements and measurement noise in the eyetracking device.

3.2.1 *Representation.* We represent the cropping window path as a piecewise B-spline that allows for a flat segment (i.e., stationary cropping window), followed by a smoothly varying segment (i.e., pan with ease-in-ease-out), followed by a flat segment (cropping window stationary again). We impose C^1 smoothness on the computed cropping window path by representing it as a piecewise nonuniform cubic B-spline with repeated knots. A nonuniform cubic B-spline is parametrized by $(m + 1)$ control points and $(m + 5)$ knots. For $m = 7$, the control points are $\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_7$, and the knots are $t_0, t_1, \ldots, t_{11}$. The first curve segment Q_3 is controlled by $(\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3)$ and defined in the interval $[t_3, t_4)$. In general, the curve segment Q_i is computed as

$$Q_i(t) = \mathbf{P}_{i-3} \cdot \mathbf{B}_{i-3,4} + \mathbf{P}_{i-2} \cdot \mathbf{B}_{i-2,4} + \mathbf{P}_{i-1} \cdot \mathbf{B}_{i-1,4} + \mathbf{P}_i \cdot \mathbf{B}_{i,4}, \quad t_i \le t < t_{i+1}. \tag{1}$$

Here, $\mathbf{B}_{i,j}$ are blending functions, and $i = 3, 4, \ldots, 7$ [Foley et al. 1996].

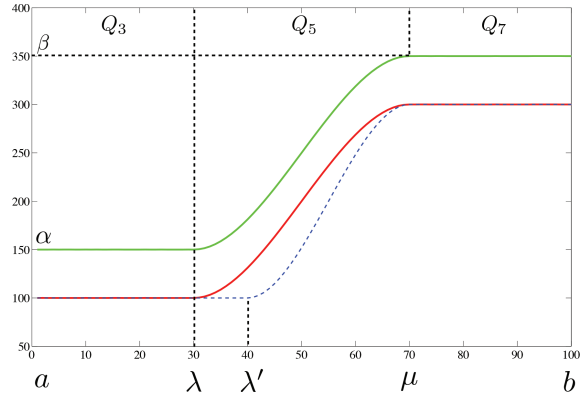

Fig. 4. Piecewise spline curves ($m = 7$) with the same knot values at endpoints ($a = 1, b = 100$) but different control points and interior knots.

By appropriately designing the multiplicity of the knots, we can influence the behavior of the curve segment. We set the control points $(\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5, \mathbf{P}_6, \mathbf{P}_7) = (\alpha, \alpha, \alpha, \alpha, \beta, \beta, \beta, \beta)$, and the knots $(t_0, t_1, \ldots, t_{11}) = (a, a, a, a, \lambda, \lambda, \mu, \mu, b, b, b, b)$. With this multiplicity, the curve segment Q_3 is controlled by $(\mathbf{P}_0, \mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3) = (\alpha, \alpha, \alpha, \alpha)$ in the interval $[t_3, t_4) = [a, \lambda)$. The segment Q_4 is of zero length because $t_4 = t_5 = \lambda$. The segment Q_5 is controlled by $(\mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4, \mathbf{P}_5) = (\alpha, \alpha, \beta, \beta)$ in the interval $[t_4, t_5) = [\lambda, \mu)$, while the segment Q_7 is controlled by $(\mathbf{P}_4, \mathbf{P}_5, \mathbf{P}_6, \mathbf{P}_7) = (\beta, \beta, \beta, \beta)$ in the interval $[t_7, t_8) = [\mu, b)$.

The value a is the start frame of the shot, and the value b is the end frame of the shot. Therefore, the free parameters are $(\alpha, \beta, \lambda, \mu)$. Figure 4 illustrates the effect of changing the parameters of the curve. The parameters for the red plot are $\alpha = 100$, $\beta = 300$, $\lambda = 30$, $\mu = 70$. By changing the control points to $\alpha = 150$, $\beta = 350$, but keeping the same knots, we can shift the curve (green). The blue dotted line shows the effect of changing a knot to $\lambda' = 40$.

3.2.2 *Fitting the Cropping Window Path to Data.* Recorded gaze data consists of noisy measurements of "what is important" to a viewer because it is subject to individual idiosyncrasies and measurement error. We fit a nonuniform cubic B-spline robustly to this data with a RANSAC algorithm [Fischler and Bolles 1981]. The trial set τ for each RANSAC iteration comprises of four randomly selected gaze points. This selection is done by picking the first four items of a random permutation of all frames, and then selecting a random gaze sample from each frame (to avoid picking two trial points from the same frame).

The goal now is to find a curve, parametrized by $(\alpha, \beta, \lambda, \mu)$, that minimizes the saliency fitting error e_s,

$$e_s = \sum_{i,j \in \tau} ||x_i - \tilde{x}_j^i||_2, \tag{2}$$

where x_i is the center of the cropping window for the i^{th} frame (obtained by sampling the curve $Q(t)$ at $t = i$), and $\tilde{x}_j^i$ is the j^{th} recorded gaze point for the i^{th} frame.

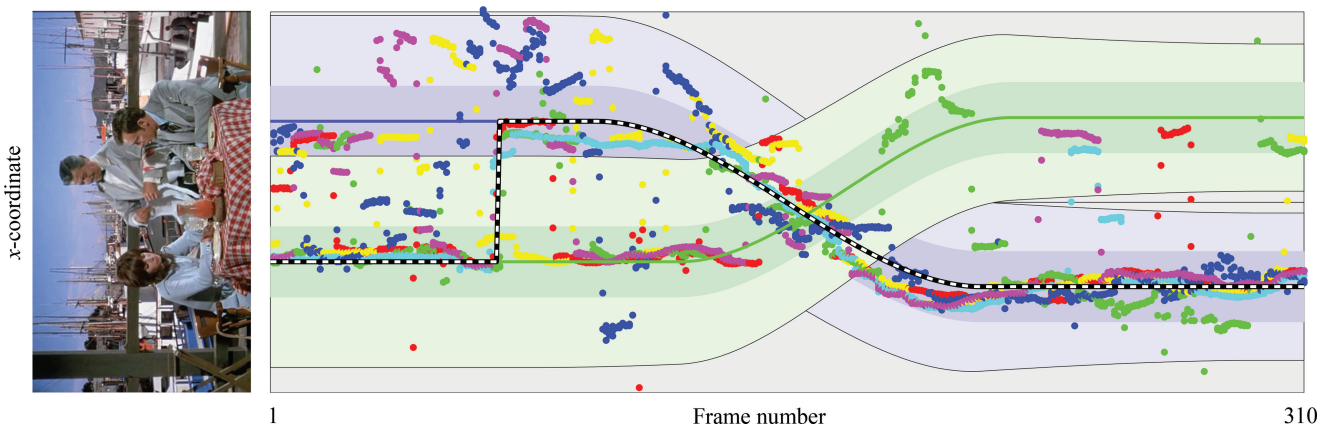

Fig. 5. The x-coordinate of the recorded gaze data is plotted for each frame of the "couple at the waterfront" video shown in Figure 1. The green and purple curves are the nonuniform B-splines computed by our algorithm corresponding to the two cropping window paths. The black dotted line is the resulting cropping window path after a cut is introduced. The gray band is the original widescreen, the light green and purple bands are the cropping windows at the given aspect ratio (1:1), and the darker green and purple bands represent the region used to compute the consensus set for each RANSAC trial. Still frame from Herbie Rides Again courtesy The Walt Disney Company.

Because the knot sequence must be nondecreasing ($a \leq \lambda < \mu \leq b$), there are three linear constraints to be satisfied:

$$a - \lambda \leq 0, \tag{3}$$
$$\lambda - \mu \leq K_1, \tag{4}$$
$$\mu \leq b, \tag{5}$$

where K_1 is a minimum distance between knots. Intuitively, K_1 limits how fast the camera is allowed to pan. Additionally, the cropping window should not exceed the boundary of the original screen, that is, $1 + \mathbf{D}_i(2)/2 \leq \alpha, \beta \leq \mathbf{D} - \mathbf{D}_i(2)/2$.

The selection of trial points is repeated N times and, for each trial, we use MATLAB's constrained minimization solver to compute the piecewise spline curve that fits the current trial samples τ. The consensus set for this curve is

$$\eta = \left\{\tilde{x}_j^i : |x_i - \tilde{x}_j^i| < K_2 \mathbf{D}_i(2)\right\}. \tag{6}$$

The score of the associated trial is $|\eta|$. The parameter K_2 is set to $1/3$ to compute the number of gaze points enclosed within the third guides of the cropping window. The curve with the highest score is selected as the cropping window path. The result of this algorithm for an example video is shown in Figure 5, where the green and purple curves are two possible pans.

3.3 Cuts

When viewer attention shifts quickly across the original widescreen video, a cut may be preferable to a fast-moving cropping window; our method allows the introduction of a cut based on recorded gaze data. We fit two spline curves ${}^A x$ and ${}^B x$ to randomly selected trial sets τ_A and τ_B.[2] A cut is generated by switching from ${}^A x$ to ${}^B x$ based on viewer attention shifts. In practice, we find that cutting more than once in a shot from a professionally edited movie is unnecessary because the shots are tightly edited to begin with.

A shift in viewer attention is computed from the change in median gaze position across consecutive frames. Candidates for cuts are posited when the shift in the median is above a threshold value and the two cropping windows A and B are sufficiently far apart (to avoid a "jump" cut, a cut that appears jarring to the viewer because the scenes before and after the cut are too similar [Dmytryk 1984]).

$$\underset{j=1\cdots\gamma_i}{\text{median}}\left(\tilde{x}_j^{i+1}\right) - \underset{j=1\cdots\gamma_i}{\text{median}}\left(\tilde{x}_j^{i}\right) > K_3, \tag{7}$$
$$\left|{}^A x^i - {}^B x^i\right| > K_4, \tag{8}$$

where γ_i is the number of gaze points recorded for frame i. Each candidate cut κ is ranked by the number of gaze points enclosed by the resulting cropping window path. Let x_i^{final} denote the cropping window path for the cut κ:

$$x_i^{\text{final}} = \begin{cases} {}^A x_i & \text{if } i \leq \kappa, \\ {}^B x_i & \text{otherwise.} \end{cases} \tag{9}$$

Then, the consensus set is computed as in Eq. (6), $\eta = \{\tilde{x}_j^i : |x_i^{\text{final}} - \tilde{x}_j^i| < K_2 \mathbf{D}_i(1)\}$ and the score of the associated cut is $|\eta|$. If the candidate cut with the highest score κ^* encloses more gaze samples than either one of the individual cropping windows, the corresponding resulting path $x^{\text{final}*}$ is the path selected for this RANSAC iteration. In Figure 5, the final path is shown as a black-and-white dotted line.

3.4 Zoom

Reducing the size of the cropping window gives the effect of "zoom" by making the scene look bigger. We compute the change in cropping window size from the spread of the gaze data. Intuitively, when the gaze samples are clustered tightly, a smaller cropping window can capture the region underneath the gaze samples.

We measure the spread of gaze data through the standard deviation σ_i of the gaze points $\tilde{x}_j^i$ for the i^{th} frame. The ratio ρ_i is computed as

$$\rho_i = \frac{\sigma_i}{\max\limits_{k=1\cdots F}(\sigma_k)}. \tag{10}$$

We fit a nonuniform piecewise B-spline, denoted by q, to the samples ρ_i with a RANSAC algorithm similar to Section 3.2.2. The consensus set η for each RANSAC trial is the set of sample

[2]For smarter sampling, τ_B is selected from among those samples outside the consensus set of ${}^A x$.

Background objects are moving.

Scene camera \ Foreground object	slight motion	significant motion
stationary	2 clips	3 clips
moving	none	4 clips

Background objects are stationary.

Scene camera \ Foreground object	slight motion	significant motion
stationary	2 clips	7 clips
moving	1 clip	3 clips

Fig. 6. We ran our algorithm on a variety of clips (12–24 seconds in duration) taken from three Hollywood films: *Herbie Rides Again*, *Who Framed Roger Rabbit*, and *The Black Hole*. The clips are categorized based on the motion of the foreground object, the scene camera, and the background objects. When one clip fits more than one category, it is counted in both categories (for example, if the scene camera was stationary for half the duration and moved in the other half).

points that are within a given distance K_5 of the spline curve,

$$\eta = \rho_i : |\rho_i - q_i| < K_5, \tag{11}$$

where $K_5 = 0.2$. The score of the associated trial is $|\eta|$. Let the curve with the highest score be q^*. Then this curve is transformed linearly (scaled and shifted)

$$q_i^{**} = K_6 q_i^* + \left(1 - \max_{k=1\cdots F}\left(q_k^*\right)\right), \tag{12}$$

where $K_6 = 0.5$. This transformation allows us to modulate the extent of the zoom effect by changing the parameter K_6. Decreasing the value causes the zoom to become less pronounced. The shift up causes the maximum window size to be equal to the maximum possible size that will fit inside the original widescreen frame. The zoom effect is then created by scaling the size of the cropping window by q_i^{**} before rendering.

4. RESULTS

We present results on 18 video clips selected to have moving and stationary background, foreground objects, and camera. Figure 6 shows the number of clips in each category. Eight clips involve a moving camera. The foreground objects and background objects both show significant movement in seven clips.

The same parameter values were used for all the examples. The number of RANSAC trials is $N = 1000$, and the parameters are $K_1 = 100$, $K_2 = 1/3$, $K_3 = K_1$. K_4 is set to allow 20% overlap between the two cropping windows A and B,

$$K_4 = \min(\mathbf{D}_i(2), \max(\mathbf{D}(2) - 1.2\mathbf{D}_i(2), \mathbf{D}_i(2)/2)).$$

K_5 is set to 0.2, and K_6 is set to 0.5 or 1. These parameters were selected based on a clip taken from each of the example videos; the algorithm was then run on the entire duration of the example videos and the results are presented in the accompanying video submission. In Figure 7, we show a frame each from two example videos at the 1:1 aspect ratio. Figure 8 shows two examples where our method zooms in, where the parameter $K_6 = 1$. Generally, this default value works well; in the accompanying video, we also show $K_6 = 0.5$ as a comparison.

A 30-second sequence takes approximately 40 minutes of computation time because the nonuniform B-spline blending functions need to be computed for each RANSAC trial. The trials could be parallelized to reduce the computation time. The runtime for our method is independent of the resolution of the video.

Fig. 7. Sample frames are shown from two example sequences. We compare with the optimized crop-and-warp method of Wang et al. [2011]. Significant distortions are introduced by their method for large-scale changes. Images from Herbie Rides Again courtesy The Walt Disney Company.

Fig. 8. Two example videos where our method zooms into the scene. Images from Herbie Rides Again courtesy The Walt Disney Company.

5. EVALUATION

We show a comparison of our method with an optimized crop-and-warp approach [Wang et al. 2011] in Figure 7. The squeezing artifact is apparent in individual frames. We evaluate our method by eye-tracking viewers watching the result videos, as well as asking viewers to indicate their subjective preference through a forced-choice questionnaire. A comparison of viewer eye movements before and after the re-editing operations tests whether viewers are absorbing the same information, and their subjective preference tells us whether they liked what they saw.

5.1 Eyetracking

We ran the evaluation on the most aggressive aspect ratio presented in this article, that is, 1:1. The video clips were displayed at their re-edited size, with a gray background, on a 19-inch screen. Participants sat 18–24 inches from the screen, and the order of presentation was randomized.

The first check we perform is the percentage of recorded gaze data on the original video that was included in the result. This check is similar to the validation performed in Chamaret and Le Meur [2008]. The included set for each frame i of a video v is

$$\psi'_{iv} = \left\{\tilde{x}_j^i : |x_i - \tilde{x}_j^i| < \mathbf{D}_i(2)\right\} \quad i = 1, \ldots, N_v. \tag{13}$$

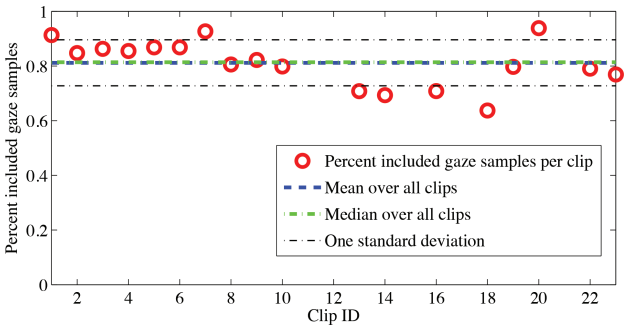

Fig. 9. The red markers show the mean percent of included gaze samples for each example clip. The mean percentage for all example videos is 81.2%($\sigma = 8.4$) and the median percentage is 81.4%.

The percentage of gaze data included for a whole video is the mean over all frames of the video, and the average percentage for V videos is

$$\psi'' = \frac{1}{V}\sum_{v=1}^{V}\sum_{i=1}^{N_v}\frac{|\psi'_{iv}|}{\gamma_i}. \tag{14}$$

For the 18 example clips we tested (12–43 seconds each), the percentage of included gaze samples are shown in Figure 9. The average included percentage is $\psi'' = 81.2\%$.

The second check examines the extent to which our re-editing algorithm alters viewer eye movements. We compare the eyetracking data of viewers watching our result videos with the eyetracking data collected on the original clips. Let $\mathbf{r}_v = [[\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n]^T, [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n]^T]$ be the gaze data on the original video, and $\mathbf{r'}_v$ be the recorded gaze data on the retargeted video, where v indexes the video clip. This data is a distribution with a per-frame mean μ, median η, and standard deviation σ.

There are several metrics to compute the distance between two distributions. A first-order metric is the distance between the per-frame mean or median values. A higher-order metric is a statistic such as the chi-squared distance or the Earth Mover's distance [Judd et al. 2012; Zhao and Koch 2012].

Because we compare viewer gaze on videos of different sizes, directly computing differences in the pixel locations of the points of regard will not yield the comparison we are looking for; we want to measure whether viewers could be looking at the same object in the result video as in the original video. Thus, the gaze data on the result video is transformed back to the coordinates of the original video with the inverse operator w.

The gaze distribution on the re-edited video will likely be different from the original distribution, even if viewers are looking at the same objects. For example, the standard deviation of the gaze data on the original widescreen video is generally larger than on the result video because the eyes move across a larger region of the screen. Therefore, metrics like chi-square and Earth Mover's distance (after histogram normalization) will return nonzero values.

In Figure 10, we plot the median gaze positions for each frame of the "lawyer" video. Then, the distance δ between included gaze data on the original video (i.e., those red markers that are inside the colored portion of the frame in Figure 10) and gaze data on the retargeted video is defined as

$$\delta = \frac{1}{V}\sum_{v=1}^{V}\Delta\big(\eta\big(w\big(\mathbf{r'}_v\big)\big), \eta\big(\mathbf{r}_v^{\text{incl}}\big)\big), \tag{15}$$

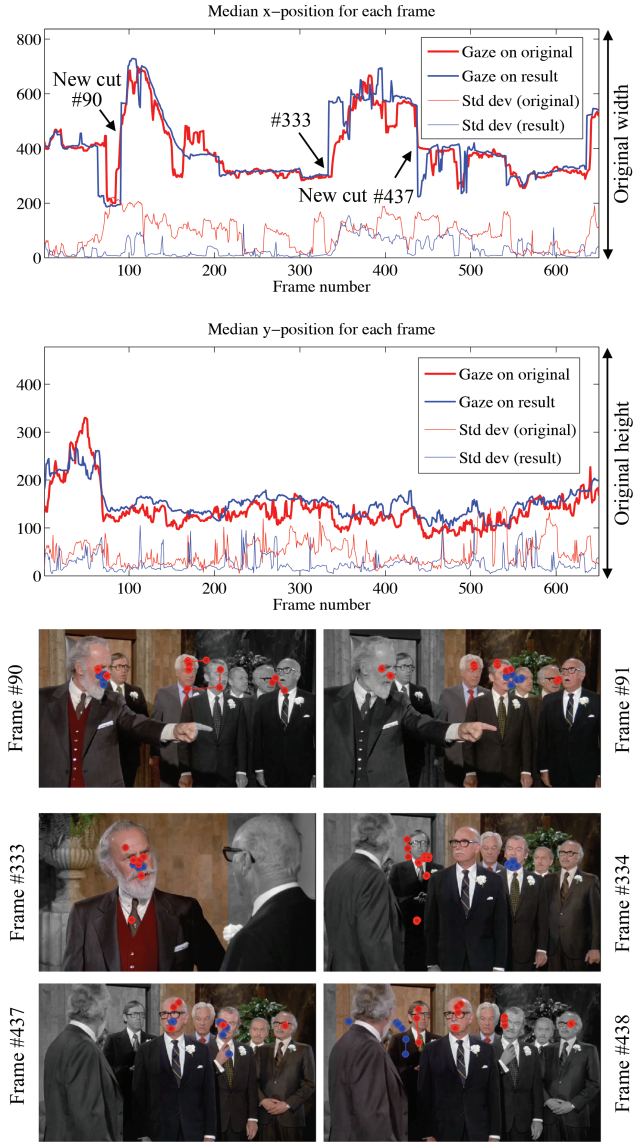

Fig. 10. The gaze data captured on the original widescreen video is shown in red and the data captured on our result (transformed to widescreen coordinates) is shown in blue. We plot the median instead of the mean to reduce the effect of outlier gaze data on the computed distance values. The thick blue line represents the median gaze data recorded on our result videos and the thick red line represents the median gaze data for the original widescreen videos. The thin blue and red lines represent the standard deviations for each frame. Images from Herbie Rides Again courtesy The Walt Disney Company.

where Δ is the root mean squared distance between the median gaze position per frame, averaged over all the frames for video v. For our 18 example clips, the distance Δ is plotted in red in Figure 11 and the average distance $\delta = 79.3$, shown as a blue dotted line. Sample frames are shown in Figure 10.

We take a closer look at two cuts introduced by our method; the corresponding frames are shown in the first and third rows in Figure 10. Our method selected the appropriate locations to

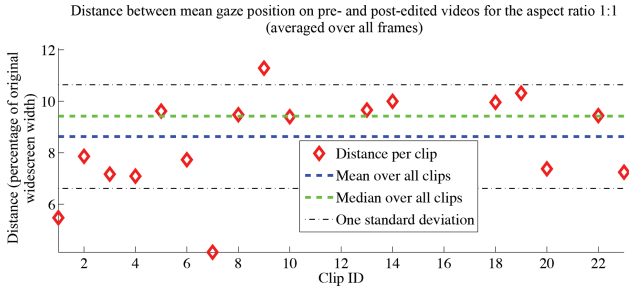

Fig. 11. The red markers show the percentage distance between the median gaze position on our results at the 1:1 aspect ratio and the median position of the included gaze samples for the original widescreen video, that is, the distance between the median of the blue markers and the median of those red markers that lie on the colored portion of the frame in Figure 10. The average distance over all example clips is 79.3 pixels ($\sigma = 21.8$), which is less than 10% of the width of the original widescreen frame.

introduce the cuts because viewer gaze shifts from left to right and right to left, respectively, in the original widescreen video. This shift can be seen in the sample frames and on the graph where we have marked the frames with the label "New cut". The sample frames shown in the second row correspond to a cut that was part of the original video (Frames #333 and #334). Our algorithm placed the cropping window to the right side of the original widescreen after this cut. It takes some time for the viewers of the original widescreen video to shift their attention to the group of lawyers.

The consistency in viewer eye movements on the original video and our re-edited result (as illustrated for the lawyer video in Figure 10) suggests that viewers are absorbing the same information from our re-edited results as from the original widescreen videos. This indicates that, even though our method crops away large parts of the screen and introduces additional camera movement, viewers are not distracted away from those regions of the video that are important for the understanding of the narrative. The eyetracking-based evaluation also reveals the importance of continuity editing. Because our method treats each shot independently, the viewer eye movements lag by a couple of frames from Frame #333 to Frame #334 in Figure 10. We discuss the use of cinematic conventions for future algorithms in Section 6.

In Figures 12 and 13, we plot two higher-order distance metrics: the chi-squared distance and the Earth Mover's distance. Both metrics histogram the gaze locations, and then compute the distance between the two histograms, thus the computed values depend on the number of bins. We report the distances for three and 10 bins, respectively. Because the number of gaze samples per frame is small (approximately 6–12) and may be different for the original widescreen video and the corresponding result video, both metrics yield small nonzero numeric values even when the gaze distributions are quite similar.

The videos showing the eyetracking data on the input and on the output are available as supplementary material. All the evaluated result videos are pans and cuts, without zoom. Gaze data does not clearly test the zoom operation because a viewer's gaze patterns are unlikely to be altered whether he or she sees a face at its original size or a little larger. Gaze data does reveal whether the viewer was led to look in a certain direction by the camera pan or missed seeing a portion of the frame because our method cut too early. Thus, we collect eyetracking data on re-edited results with pans and cuts.

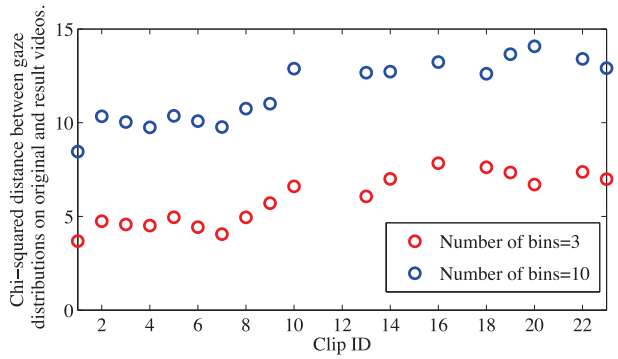

Fig. 12. We compute the chi-squared distance between the gaze data on the original widescreen video and our result for each frame, and we report the average distance over all frames in a clip. Because this distance metric depends on the number of bins in the underlying histogram, we show the values for three bins in red and 10 bins in blue. We show results on 18 videos from a database of 23 clips. Clip ID refers to the index of the video in the database.

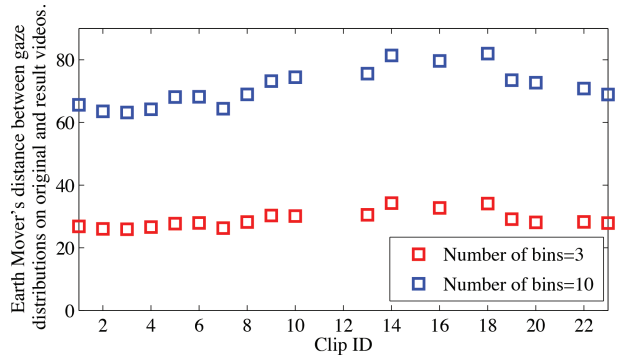

Fig. 13. We compute the Earth Mover's distance between the gaze data on the original widescreen video and our result for each frame, and report the average distance over all frames in a clip. Because this distance metric depends on the number of bins in the underlying histogram, we show the values for three bins in red and 10 bins in blue. Even though the gaze distributions are quite similar, the distance metric does not return a zero value because the total number of gaze points can be different.

5.2 User Preference

We also evaluate our method by asking users to submit their preferences. We conduct a two-alternative forced-choice study where 15 participants compared results generated by our method with those generated by an optimized crop-and-warp method by Wang et al. [2011], and the letterboxed version of the same clip.

We use five video clips in our experiment. The participants first watch the original widescreen clip, then one version of this clip at the reduced aspect ratio 1:1 ("Video 1"), and then another version of the clip at the reduced aspect ratio ("Video 2"). After this, they are required to answer the question "which is a better representation of the original clip at this reduced aspect ratio? (Video 1 or Video 2)". Because the order of presentation could bias the participants, we repeat the comparison with the order switched, thus each participant rates 5 videos $\times$ 2 blocks $=$ 10 video comparisons of our method versus Wang et al. [2011], and similarly for our method versus letterboxing. The order of presentation of the blocks is counterbalanced across participants.

Table I.
Preferences of the 14 participants in our subjective user evaluation at full size. Our re-editing approach (GDR) was preferred 84% over a state-of-the-art nonlinear crop-and-warp method (W2011). Letterboxing was preferred for 85% of the comparisons with GDR.

↓ preferred over →	GDR	W2011	Letterboxing
GDR	-	84%	15%
W2011	16%	-	-
Letterboxing	85%	-	-

Table II.
Preferences of the 25 participants in our subjective user evaluation at small size. Our re-editing approach (GDR) was preferred in 62% of the comparisons with a crop-and-warp method (W2011). Letterboxing was preferred in 75% of the comparisons with GDR.

↓ preferred over →	GDR	W2011	Letterboxing
GDR	-	62%	25%
W2011	38%	-	-
Letterboxing	75%	-	-

Table III.
Order effects in preferences of 25 participants in our subjective user evaluation at small size. Because there is no clear trend whether the subjects always prefer our results when it is shown first, we can conclude that repeating the comparison in both orders is effective at mitigating bias due to presentation order.

GDR preferred over	"Video 1" is GDR	"Video 2" is GDR
W2011	63%	61%
Letterboxing	23%	26%

Previously, comparisons have been designed so that the participant views the original video and the two retargeted versions side by side. This design is useful to find out whether visual artifacts are noticeable to viewers, because a side-by-side comparison encourages the viewer to distinguish, the two videos. If a viewer discerns our result from a method based on nonlinear rescaling, our method would be rated as more preferable because it does not introduce the artifacts that rescaling does (structural artifacts such as waving or squeezing). We design our study to present each video sequentially. This presentation allows the viewer to assess each clip individually, undistracted by a second clip playing simultaneously.

We perform our evaluation on clips long enough to contain a context or a conversation to better mimic the experience of watching an actual film. The five clips are 12, 9, 15, 23, and 6 seconds in duration, respectively. The complete experiment takes about 17–20 minutes to complete. We also choose to administer the experiment in situ to be able to control the quality of video playback. Participants were recruited through a Web site in accordance with IRB protocol. The participants watched the videos while sitting 18–24 inches from on a 19-inch screen. We ran a total of 15 participants (age range 18–55 years). We discarded the data of one participant as she was clearly distracted during the experiment.

The data collected from the experiment is shown in Table I. Each participant compared our method (GDR) to Wang et al. [2011] (W2011) 10 times, leading to a total of 140 comparisons. Of the 10 comparisons a participant rated, the average number of times a participant preferred GDR over W2011 is $\mu = 8.42(\sigma = 2.59)$. This observation is significantly greater than the chance value $\mu_{chance} = 5$, based on a one-tailed t-test, with $p < 0.001, t(13) = 4.95$. The effect size of $r = 0.8081$ indicates a large effect. The observed statistical power is 0.998 [Erdfelder et al. 1996], suggesting that even a small number of participants is sufficient to observe the effect.

As shown in Table I, each participant compared our method (GDR) to letterboxing 10 times, leading to a total of 140 comparisons. On average, out of the 10 comparisons a participant rated, he or she preferred letterboxing to GDR $\mu = 8.5$ times ($\sigma = 1.56$). The preference values are significantly greater than chance ($\mu_{chance} = 5$), based on a one-tailed t-test, with $p < 0.001$, $t(13) = 8.4130, r = 0.9191$. The observed statistical power is 0.99.

5.3 User Preference at Mobile Phone Size

We additionally performed an evaluation of our method at a viewing size that mimics a mobile phone. Because mobile phones range in size from 2.3–2.8 inches, we present the videos at 2.6 inches in height when shown on a 15inch laptop screen at its native resolution (1440×900). The background was set to gray. The experiment design was identical to that in Section 5.2. We recruited 25 naive participants through a Web site in accordance with IRB guidelines, leading to a total of 250 preference values for each comparison. The collected data are summarized in Table II.

The pixels per degree of visual angle (ppd) were comparable for the "regular-size" experiment and the "mobile-phone-sized" experiment, and the values were comparable to the pixels per degree for mobile phones (ranging from 34–46 ppd for external monitor, 28–46 ppd for laptop screen, and 30–53 ppd for mobile phones). Because visual resolution comparisons are meaningful only for normal viewing distances for the device considered, mobile phones and laptops or desktops have very different pixel per-inch resolutions, but similar pixel per-degree resolutions. As the variety of display devices increases, perception researchers are beginning to study how the interaction of pixel resolution and visual angle impacts people's abilities to resolve visual properties [Healey and Sawant 2012].

We found that, on average, out of 10 comparisons, a participant preferred GDR over W2011 $\mu = 6.24$ times ($\sigma = 2.88$). This preference is greater than the chance preference value $\mu_{chance} = 5$ based on a one-tailed t-test ($p = 0.02$, $p < 0.05, t(24) = 2.15$). The effect size is $r = 0.403$ and post-hoc statistical power is 0.67. Additionally, participants on average preferred letterboxing to GDR $\mu = 7.52$ times ($\sigma = 2.92$). This observation is highly significant ($p < 0.001, t(24) = 4.31$) based on a one-tailed t-test. The effect size is $r = 0.66$ and the observed power is 0.99. These findings indicate that user preferences at mobile phone sizes follow the same trend: our results on gaze-driven video re-editing are preferred relative to an optimized crop-and-warp method, and letterboxed videos are preferred relative to the re-edited videos.

6. DISCUSSION

We have presented an algorithm for re-editing videos to better fit a device with a smaller aspect ratio. Our method uses the gaze data of viewers on the original video to determine the important parts of the frame. We compute B-spline paths for two cropping windows, and find an optimal cut between them using a RANSAC approach. The zoom is computed from the per-frame standard deviation of the input gaze data, and we compare the gaze data of viewers on the re-edited videos with that on the original videos. The median location of viewer gaze on the pre- and post-edited videos is within 10% of the frame dimension.

We evaluate our method through a two-alternative forced-choice user study that asks viewers to indicate their preference for one of two videos. The results show that viewers find our results to be a

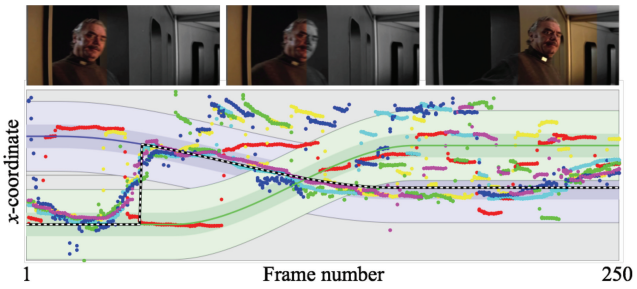

Fig. 14. Top row: Our result in color is overlaid on the grayscale original frame. Because the character darts quickly across the frame, the required pan velocity is too high and our algorithm introduces a cut instead. This creates a result with a "jump cut". Bottom row: Viewer gaze for this video and the cropping window path computed by our method. See Figure 5 for an explanation of the colors. Images from The Black Hole courtesy The Walt Disney Company.

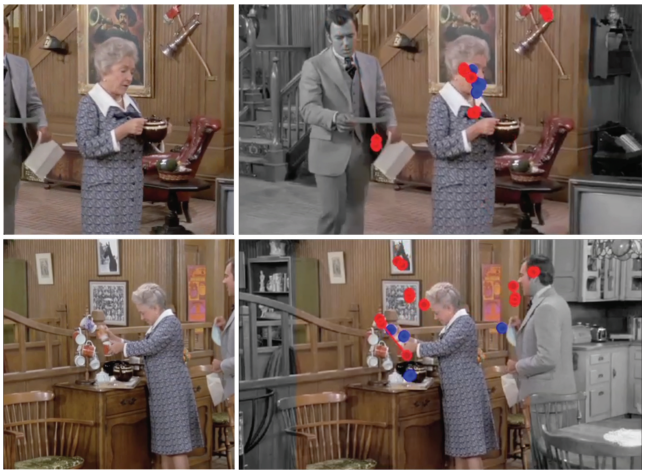

Fig. 15. Left column: Our result. Right column: Our result in color is overlaid on the grayscale original frame. Red dots mark viewer gaze on the original video, and blue dots mark viewer gaze on the re-edited result. Because eyetracking data only indicates the pixels that are being attended, it is possible for a cropping window to truncate a secondary character while optimizing the smoothness of the pan. Though our comparison of viewer gaze on pre- and post-edited videos shows that these truncations do not distract attention away from the primary character, a human editor might select a different cropping window to avoid this artifact. Images from Herbie Rides Again courtesy The Walt Disney Company.

better representation of the original video compared to an optimized crop-and-warp method, and letterboxed videos to be a better representation of the original video compared to our method. This study was performed at two sizes: a "regular" size and a "mobile phone" size. The finding that viewers prefer watching the video at the original aspect ratio (even after letterboxing) perhaps reflects audience validation of the directorial vision that caused a certain aspect ratio to be chosen for a film (for example, Sydney Pollack's lawsuit for his 1975 film [Young 2008]). Alternatively, this finding could be a result of a bias introduced by the question of what constitutes a "better" representation of the original video.

We also ran an informal study where viewers were shown videos resized to a mobile phone size by our method, crop-and-warp, and letterboxing, and asked to rate each video on a 7-point scale. We found a significant difference in the responses to the questions: "How well are you able to see the expressions on the actors' faces?" and "How much would you want to watch this video on a mobile phone?" For both questions, our method was rated better than either letterboxing or crop-and-warp. Interestingly, there was no significant difference in the ratings for the question, "How well can you see the action", indicating perhaps that close-ups are less required for conveying that the actor is moving across the screen. These trends suggest that the standard two-alternative forced-choice comparison metric for video retargeting methods might not capture all the nuances of viewing experiences at differently sized devices. An exploration of the evaluation metrics used by the community would make for an interesting and useful future direction.

A limitation of our algorithm is the extent to which it can re-edit a scene with fast motion or short cuts. When the algorithm fails, it is most often because the reduction that we have asked for is impossible without two or more cuts or a high-velocity pan. Figure 14 illustrates the performance of our method on a scene with fast motion where the character darts out of the door and through the corridor. Because of the threshold on how fast a single cropping window is allowed to pan is kept the same for all our examples, our method selects a cut as the optimal solution. This solution results in a "jump cut". Another limitation of our method is that it is applied to individual shots, which can lead to sudden camera moves during gradual shot transitions such as dissolves. This type of artifact could be addressed by introducing a continuity term for edits across shots.

As eyetracking technologies become cheaper and more easily available (for example, Webcam-based eyetracking [Agustin et al. 2010; Abbot and Aldo 2011]), it will become possible to obtain eyetracking information even by crowdsourcing viewer gaze [Rudoy et al. 2012]. Hence, collecting large quantities of gaze data will become feasible in terms of both quality and cost. In the future, eyetracking could also be done on smart personal devices such as smartphones, and on a per-user basis to automatically create a personalized version of a movie. Such research would provide a big data complement to current work on understanding how viewers view television and film on mobile devices (for example, Knoche et al. [2008]).

For viewer attention to be a viable input to computer graphics algorithms, we need an estimate for how many viewers should be eyetracked to access the "canonical" eye movement pattern. We computed the change in mean gaze location per frame of a video clip as the number of viewers in the database was increased. The shift in mean gaze location levels-off at around four or five viewers. The numeric value will not reach zero because it is not necessary for the true distribution of gaze locations to be drawn from a single Gaussian distribution; it could be a mixture distribution if the scene included two people talking and both the expression of the speaker and the reaction of the listener were salient.

Our re-editing algorithm guarantees that the structure of the underlying scene will be preserved, that is, faces will not be made tall and skinny and lines will not be broken. However, a drastic change in aspect ratio will necessarily compel the re-editing algorithm to crop away large parts of the underlying scene, thus it might seem that our algorithm is susceptible to artifacts such as cutting faces in half. However, as the re-edit is driven by the viewer's gaze, it avoids cutting off a region that was attended to, and secondary faces or limbs may get truncated (for example, Figure 15). Our comparison of viewer gaze on pre- and post-edited videos shows that, because the regions attended by the viewer on the original video are preserved during the re-edit, the secondary faces or limbs that may get truncated do not alter the viewer's attention patterns.

Our algorithm for placing cuts is largely successful in inserting them so that they are not disruptive to the viewing pattern. However, our approach does tend towards using cuts because two windows will always provide more coverage of the salient regions than one, particularly with the relatively slow panning velocities that we have allowed. A statistical analysis of the camera motion and editing patterns used in the original film might help to provide additional information about when difficult scenes could be re-edited with more cuts, faster pans, or zooms. Eye movement research indicates that cinematic guidelines such as continuity editing result in viewers experiencing increased "edit blindness" [Smith and Henderson 2008]. Continuity editing rules include techniques such as matching the action before and after the cut, and aligning a cut to a sound effect. Future work could incorporate such guidelines from the study of cinema (for example, Dmytryk [1984]) to create more pleasing re-edits. Combining such an analysis with gaze data might also benefit other applications such as reframing for aesthetic reasons (see, e.g., Liu et al. [2010]), and video summarization.

REFERENCES

W. Abbot and F. Aldo. 2011. Ultra-low cost eyetracking as an high information throughput alternative to BMIS. *BMC Neurosci.* 12,1.

J. S. Agustin, H. Skovsgaard, E. Mollenbach, M. Barret, M. Tall, D. W. Hansen, and J. P. Hansen. 2010. Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the Symposium on Eyetracking Research and Applications (ETRA'10).* 77—80.

S. Avidan and A. Shamir. 2007. Seam carving for content-aware image resizing. *ACM Trans. Graph.* 26, 3.

F. Baluch and L. Itti. 2011. Mechanisms of top-down attention. *Trends Neurosci.* 34, 210–224.

S. Castillo, T. Judd, and D. Gutierrez. 2011. Using eye-tracking to assess different image retargeting methods. In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV'11).*

C. Chamaret and O. Le Meur. 2008. Attention-based video reframing: Validation using eye-tracking. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08).*

D. DeCarlo and A. Santella. 2002. Stylization and abstraction of photographs. *ACM Trans. Graph.* 21, 3, 769–776.

T. Deselaers, P. Dreuw, and H. Ney. 2008. Pan, zoom, scan – Time coherent, trained automatic video cropping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08).* 1–8.

E. Dmytryk. 1984. *On Film Editing*. Focal Press.

M. Dorr, T. Martinetz, K. Gegenfurtner, and E. Barth. 2010. Variability of eye movements when viewing dynamic natural scenes. *J. Vis.* 10, 10.

H. El-Alfy, D. Jacobs, and L. Davis. 2007. Multi-scale video cropping. In *Proceedings of the 15th ACM International Conference on Multimedia (MULTIMEDIA'07).* 97–106.

E. Erdfelder, F. Faul, and A. Buchner. 1996. Gpower: A general power analysis program. *Behav. Res. Meth. Instrum. Comput.* 28, 1, 1–11.

M. A. Fischler and R. C. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM* 24, 381–395.

J. D. Foley, A. Van Dam, S. K. Feiner, and J. F. Hughes. 1996. *Computer Graphics Principles and Practice* 2nd Ed. Addison-Wesley.

R. B. Goldstein, R. L. Woods, and E. Peli. 2007. Where people look when watching movies: Do all viewers look at the same place? *Comput. Biol. Med.* 37, 7, 957–964.

C. G. Healey and A. P. Sawant. 2012. On the limits of resolution and visual angle in visualization. *ACM Trans. Appl. Percept.* 9, 4, 20:1–20:21.

E. Jain, Y. Sheikh, and J. Hodgins. 2012. Inferring artistic intention in comic art through viewer gaze. In *Proceedings of the ACM Symposium on Applied Perception (SAP'12).*

T. Judd, F. Durand, and A. Torralba. 2012. A benchmark of computational models of saliency to predict human fixations. Tech. rep. MITCSAIL-TR-2012-001, Massachusetts Institute of Technology. http://dspace.mit.edu/handle/1721.1/68590.

H. Katti, A. K. Rajagopal, M. Kankanhalli, and R. Kalpathi. 2014. Online estimation of evolving human visual interest. *ACM Trans. Multimedia Comput. Comm. Appl.* 11, 1.

S. D. Katz. 1991. *Shot by Shot*. Michael Wiese Productions, Focal Press.

H. Knoche, J. McCarthy, and M. Sasse. 2008. How low can you go? The effect of low resolutions on shot types in mobile tv. *Multimedia Tools Appl.* 36, 1–2, 145–166.

S. Kopf, T. Haenselmann, J. Kiess, B. Guthier, and W. Effelsberg. 2011. Algorithms for video retargeting. *Multimedia Tools Appl.* 51, 2, 819–861.

P. Krähenbühl, M. Lang, A. Hornung, and M. Gross. 2009. A system for retargeting of streaming video. *ACM Trans. Graph.* 28, 126:1–126:10.

F. Liu and M. Gleicher. 2006. Video retargeting: Automating pan and scan. In *Proceedings of the ACM International Conference on Multimedia (MULTIMEDIA'06).* 241–250.

L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. 2010. Optimizing photo composition. *Comput. Graph. Forum* 29, 2, 469–478.

P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. 2010. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cogn. Comput.* 3, 1, 5–24.

Y. Niu, F. Liu, X. Li, and M. Gleicher. 2010. Warp propagation for video resizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10).* 537–544.

M. Rubinstein, A. Shamir, and S. Avidan. 2008. Improved seam carving for video retargeting. *ACM Trans. Graph.* 27, 3, 16:1–16:9.

D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. 2012. Crowdsourcing gaze data collection. http://arxiv.org/abs/1204.3367.

A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'06).* 771–780.

A. Shamir and O. Sorkine. 2009. Visual media retargeting. In *Proceedings of the 1st ACM SIGGRAPH Conference and Exhibition in Asia (SIGGRAPH-ASIA'09).* 11:1–11:13.

T. J. Smith and J. M. Henderson. 2008. Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *J. Eye Movement Res.* 2, 2, 1–17.

C. Tao, J. Jia, and H. Sun. 2007. Active window oriented dynamic video retargeting. In *Proceedings of the Workshop on Dynamical Vision at the International Conference on Computer Vision (ICCV'07).*

J. Wang, M. J. T. Reinders, R. L. Lagendijk, J. Lindenberg, and M. S. Kankanhalli. 2004. Video content representation on tiny devices. In *Proceedings of the IEEE Conference on Multimedia and Expo (ICME'04).* 1711–1714.

Y.-S. Wang, H. Fu, O. Sorkine, T.-Y. Lee, and H.-P. Seidel. 2009. Motion-aware temporal coherence for video resizing. *ACM Trans. Graph.* 28, 127:1–127:10.

Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee. 2011. Scalable and coherent video resizing with per-frame optimization. *ACM Trans. Graph.* 30, 4, 88:1–88:8.

Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee. 2010. Motionbased video retargeting with optimized crop-and-warp. *ACM Trans. Graph.* 29, 90:1–90:9.

Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee. 2008. Optimized scale-and-stretch for image resizing. *ACM Trans. Graph.* 27, 118:1–118:8.

Wikipedia. 2015. http://en.wikipedia.org/wiki/pan_and_scan.

Y. Y. Xiang and M. S. Kankanhalli. 2010a. Automated aesthetic enhancement of videos. In *Proceedings of the ACM International Conference on Multimedia (MM'10).* 218–290.

Y.-Y. Xiang and M. S. Kankanhalli. 2010b. Video retargeting for aesthetic enhancement. In *Proceedings of the ACM International Conference on Multimedia (MM'10).* 919–922.

J. Young. 2008. Sydney Pollack dies at 73. *Variety*, May 26.

Q. Zhao and C. Koch. 2012. Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *J. Vis.* 12, 6.

Received December 2012; revised September 2014; accepted September 2014