Efficient Scale-space Spatiotemporal Saliency Tracking for Distortion-Free Video Retargeting

Gang Hua, Cha Zhang, Zicheng Liu, Zhengyou Zhang and Ying Shan

July 16, 2009

Technical Report
MSR-TR-2009-87
(A shorter version appears in the *Proc. Asian Conference on Computer Vision* (ACCV), Xi'an, China, Sept. 23-27, 2009)

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

{ganghua,chazhang,zliu,zhang,yingsh}@microsoft.com
http://research.microsoft.com/~zhang

Efficient Scale-space Spatiotemporal Saliency Tracking for Distortion-Free Video Retargeting

Gang Hua, Cha Zhang, Zicheng Liu, Zhengyou Zhang and Ying Shan

Microsoft, One Microsoft Way, Redmond, WA 98052, USA
{ganghua, chazhang, zliu, zhang, yingsh}@microsoft.com

Abstract. Video retargeting aims at transforming an existing video in order to display it appropriately on a target device, often in a lower resolution, such as a mobile phone. To preserve a viewer's experience, it is desired to keep the important regions in their original aspect ratio, i.e., to maintain them distortion-free. Most previous methods are susceptible to geometric distortions due to the anisotropic manipulation of image pixels. In this paper, we propose a novel approach to distortion-free video retargeting by scale-space spatiotemporal saliency tracking. An optimal source cropping window with the target aspect ratio is smoothly tracked over time, and then isotropically resized to the retargeted display. The problem is cast as the task of finding the most spatiotemporally salient cropping window with minimal information loss due to resizing. We conduct the spatiotemporal saliency analysis in scale-space to better account for the effect of resizing. By leveraging integral images, we develop an efficient coarse-to-fine solution that combines exhaustive coarse and gradient-based fine search, which we term scale-space spatiotemporal saliency tracking. Experiments on real-world videos and our user study demonstrate the efficacy of the proposed approach.

1 Introduction

Video retargeting aims at modifying an existing video in order to display it appropriately on a target display of different size and/or different aspect ratio [1–3]. The vast majority of the videos captured today have 320×240 pixels or higher resolutions and standard aspect ratio 4:3 or 16:9. In contrast, many mobile displays have low resolution and non-standard aspect ratios. Retargeting is hence essential to video display on these mobile devices. Recently, video retargeting has been applied in a number of emerging applications such as mobile visual media browsing [3–6], automated lecture services [7], intelligent video editing [8, 9], and virtual directors [10, 7].

In this work, we focus on video retargeting toward a smaller display, such as that of a mobile phone. Directly resizing a video to the small display may not be desirable, since by doing so we may either distort the video scene, which is visually disturbing, or pad black bars surrounding the resized video, which wastes precious display resources. To bring the best visual experiences to the users, a good retargeted video should preserve as much the visual content in the original video as possible, and it should ideally be distortion-free. To achieve this goal, we need to address two important problems: 1) how to quantify the importance of visual content? 2) How to preserve the visual content while ensuring distortion-free retargeting?

Fig. 1. Retargeting system overview: scale-space spatiotemporal saliency map (b) is calculated from consecutive n video frames (a). A minimal information loss cropping window with the target aspect ratio is identified via smooth saliency tracking (c), and the cropped image (d) is isotropically scaled to the target display (e). This example retargets 352×288 images to 100×90.

Previous works [11, 12, 1, 4, 2] approach to the first problem above by combining multiple visual cues such as image gradient, optical flow, face and text detection results etc. in an ad hoc manner to represent the amount of content information at each pixel location (a.k.a. the saliency map). It is desirable to have a simple, generic and principled approach to accounting for all these different visual information. In this paper, we improve and extend the spectrum residue method for saliency detection in [13] to incorporate temporal and scale-space information, and thereby obtain a *scale-space spatiotemporal saliency map* to represent the importance of visual content.

Given the saliency map, retargeting should preemptively preserve as many salient image pixels as possible. Liu and Gleicher [1] achieve this by identifying a cropping window which contains the most visual salient pixels and then anisotropically scale it down to fit with the retargeting display (i.e., allowing different scaling in horizontal and vertical directions). The cropping window is restricted to be of fixed size within one shot, and the motion of the cropping window can only be one of the three types, i.e., static, horizontal pan, or a virtual cut. It can not perform online live retargeting since the optimization must be performed at the shot level. Avidan and Shamir [11] use dynamical programming to identify the best pixel paths to perform recursive cut or interpolation for image resizing. Wolf et. al [2] solve for a saliency aware global warping of the source image to the target display size, and then resample the warped image to the target size. Nevertheless, it is not uncommon for all the aforementioned methods to introduce geometry distortions to the video objects due to the anisotropic manipulation of the image pixels.

In this paper, we propose to smoothly track an optimal cropping window with the target aspect ratio across time, and then isotropically resize it to fit with the target display. Our approach is able to perform online retargeting. We propose an efficient coarse-to-fine search method, which combines coarse exhaustive search and gradient based fine search, to track an optimal cropping window over time. Moreover, we only allow isotropic scaling during retargeting, and therefore guarantee that the retargeted video is distortion-free. An overview of our retargeting system is presented in Fig. 1.

There are two types of information loss in the proposed retargeting process. First, when some regions are excluded due to cropping, the information that they convey are lost. We term this the *cropping information loss*. Second, when the cropped image is

scaled down, details in the high frequency components are thrown away due to the low pass filtering. This second type of loss is called the *resizing information loss*. One may always choose the largest possible cropping window, which induces the smallest cropping information loss, but may potentially incur huge amount of resizing information loss. On the other hand, one can also crop with exactly the target display size, which is free of resizing information loss, but may result in enormous cropping information loss. Our formulation takes both of them into consideration and seeks for a trade-off between the two. An important difference between our work and [1] is that the resizing information loss we introduce is *content dependent*, which is based on the general observation that some images may be downsized much more than some other images without significantly degrading their visual quality. This is superior to the naive content independent scale penalty (a cubic loss function) adopted in [1].

The main contributions of this paper therefore reside in three-fold: **1)** we propose a distortion-free formulation for video retargeting, which yields to a problem of *scale-space spatiotemporal saliency tracking*. **2)** By leveraging integral images, we develop an efficient solution to the optimization problem, which combines a coarse exhaustive search and a novel gradient-based fine search for scale-space spatiotemporal saliency tracking. **3)** We propose a computational approach to scale-space spatiotemporal saliency detection by joint frequency, scale space, and spatiotemporal analysis.

The remainder of the paper is organized as follows. A general distortion-free video retargeting framework is introduced in Sec. 2. Salient region detection and tracking is presented in Sec. 3. The novel scale-space spatiotemporal saliency computation is illustrated in Sec. 4. Experimental results and conclusions are presented in Sec. 5 and 6, respectively.

2 Distortion-Free Video Retargeting

2.1 Problem Formulation

Consider an original video sequence with T frames $\mathcal{V} = \{I_t, t = 1, \cdots, T\}$. Each frame is an image array of pixels $I_t = \{I_t(i,j), 0 \leq i < W_0, 0 \leq j < H_0\}$, where W_0 and H_0 are the width and height of the images. For retargeting, the original video has to be fit into a new display of size $W_r \times H_r$. We assume $W_r \leq W_0, H_r \leq H_0$.

To ensure that there is no distortion during retargeting, we allow only two operations on the video – cropping and isotropic scaling. Let $\mathcal{W} = \{(x,y),(W,H)\}$ be a rectangle region in the image coordinate system, where (x,y) is the top-left corner, and W and H are the width and the height. The cropping operation on frame I_t can be defined as $\mathcal{C}_{\mathcal{W}}(I_t) \triangleq \{I_t(m+x, n+y), 0 \leq m < W, 0 \leq n < H\}$, where m and n are the pixel index of the output image. The isotropic scaling operation is parameterized with a single scalar variable s (for scaling down, $1.0 \leq s \leq s_{max}$), i.e., $\mathcal{S}_s(I_t) \triangleq \{I_t(s \cdot m, s \cdot n), s \cdot m < W_0, s \cdot n < H_0\}$. Distortion-free video retargeting can be represented as a composite of these two operations on all the video frames such that $\hat{I}_t(s_t, x_t, y_t) = \mathcal{S}_{s_t}(\mathcal{C}_{\mathcal{W}_t}(I_t)), t = 1, \cdots, T$, where $\mathcal{W}_t = \{(x_t, y_t), (s_t W_r, s_t H_r)\}$ is the cropping window at frame I_t. We further denote $\hat{\mathcal{V}} = \{\hat{I}_t, t = 1, \cdots, T\}$ to be the retargeted video, and $\mathcal{P} \triangleq \{(s_t, x_t, y_t), t = 1, \cdots, T\}$ to be the set of unknown scaling

and cropping parameters, where $\mathcal{P} \in \mathfrak{R} = \{s_t, x_t, y_t | 1.0 \leq s_t \leq s_{max}, 0 \leq x_t < W_0 - s_t W_r, 0 \leq y_t < H_0 - s_t H_r\}$.

Both cropping and scaling will lead to information loss from the original video. We propose to exploit the information loss with respect to the original video as the cost function for retargeting, i.e.:

$$\mathcal{P}^* = \arg\max_{\mathcal{P} \in \mathfrak{R}} \mathbf{L}(\mathcal{V}, \hat{\mathcal{V}}), \tag{1}$$

where $\mathbf{L}(\mathcal{V}, \hat{\mathcal{V}})$ is the information loss function, which shall be detailed in Sec. 2.2. Since ensuring the smooth transition of the cropping and resizing parameters is essential to the visual quality of the retargeted video, we also introduce a few motion constraints that shall be included when optimizing Eq. (1) in Sec. 2.3.

2.2 Video Information Loss

The *cropping* and *resizing* information loss are caused by very different reasons, hence they can be computed independently. We represent the video information loss function with two terms, i.e.,

$$\mathbf{L}(\mathcal{V}, \hat{\mathcal{V}}) = \mathbf{L}_c(\mathcal{V}, \hat{\mathcal{V}}) + \lambda \mathbf{L}_r(\mathcal{V}, \hat{\mathcal{V}}), \tag{2}$$

where λ is the control parameter to obtain a tradeoff between the cropping information loss $\mathbf{L}_c$ and the resizing information loss $\mathbf{L}_r$, which are detailed as follows.

Cropping information loss We compute the cropping information loss based on spatiotemporal saliency maps. We assume in this section such a saliency map is available (see Sec. 4 for our computation model for the spatiotemporal saliency map).

For frame I_t, we denote the per-pixel saliency map as $\{S_t(i,j), 0 \leq i < W_0, 0 \leq j < H_0\}$. Without loss of generality, we assume that the saliency map is normalized such that $\sum_{ij} S_t(i,j) = 1$. Given $\mathcal{W}_t$, the cropping information loss at time instant t is defined as the summation of the saliency values of those pixels left outside the cropping window, i.e.,

$$\mathbf{L}_c(\mathcal{W}_t) = 1 - \sum_{(i,j) \in \mathcal{W}_t} S_t(i,j). \tag{3}$$

The cropping information loss between the original video and the retargeted video is thereby defined as $\mathbf{L}_c(\mathcal{V}, \hat{\mathcal{V}}) = \sum_{t=1}^{T} \mathbf{L}_c(\mathcal{W}_t) = T - \sum_{t=1}^{T} \sum_{(i,j) \in \mathcal{W}_t} S_t(i,j)$.

Resizing information loss The resizing information loss $\mathbf{L}_r(\mathcal{V}, \hat{\mathcal{V}})$ measures the amount of details lost during scaling, where low-pass filtering is necessary in order to avoid aliasing in the down-sampled images. For a given frame I_t, the larger the scaling factor s_t, the more aggressive the low-pass filter has to be, and the more details will be lost due to scaling. In the current framework, the low-pass filtered image is computed as $I_{s_t} = \mathcal{G}_{\sigma(s_t)}(I_t)$, where $\mathcal{G}_\sigma(\cdot)$ is a 2D Gaussian low-pass filter with isotropic covariance σ, which is a function of the scaling factor s_t, i.e., $\sigma(s_t) = \log_2(s_t), 1.0 \leq s_t \leq s_{max}$.

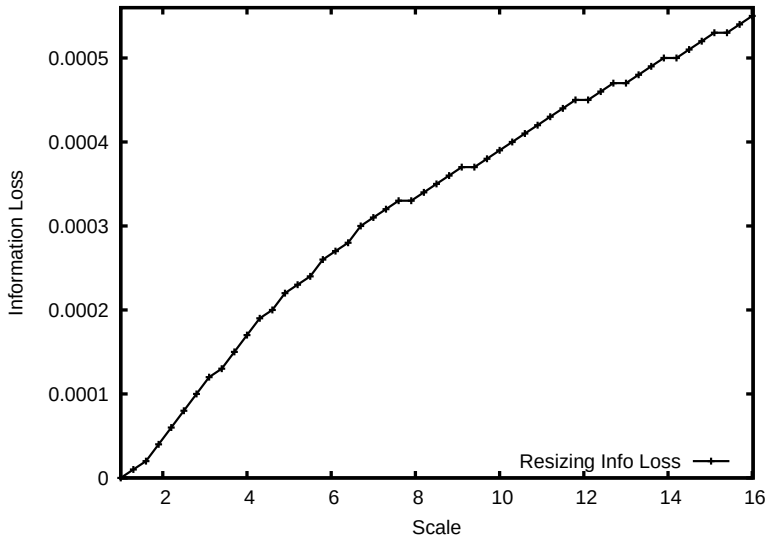

Fig. 2. Resizing information loss curve.

The resizing information loss is defined as the *squared error* between the cropped image in the original resolution and its low-pass filtered image before down-sampling, i.e.,

$$\mathbf{L}_r(\mathcal{W}_t) = \sum_{(i,j)\in\mathcal{W}_t} (I_t(i,j) - I_{s_t}(i,j))^2. \tag{4}$$

The image pixel values are normalized to be in $[0, 1]$ beforehand. For the whole video sequence, we have $\mathbf{L}_r(\mathcal{V}, \hat{\mathcal{V}}) = \sum_{t=1}^{T} \mathbf{L}_r(\mathcal{W}_t) = \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{W}_t} (I_t(i,j) - I_{s_t}(i,j))^2$. Fig. 2 presents the resizing information loss curve calculated for the cropping window presented in Fig. 1(c) using Eq. 4. As we expected, the loss function increases monotonically with the increase of the scaling factor.

2.3 Constraints for video retargeting

If there is no other additional cross-time constraints, Eq. 1 can indeed be optimized frame by frame. However, motion smoothness constraints of the cropping window, for both scaling and translation, is very important to produce visually pleasant retargeted video. To ease the optimization, we do not model motion constraints directly in our cost function. Instead we pose additional smoothness constraints on the solution space of $\mathcal{P}$ at each time instant t, i.e., the optimal $\mathcal{W}_t$ is constrained by the optimal solutions of $\mathcal{W}_{t-1}$ and $\mathcal{W}_{t-2}$. By doing so, an additional benefit is that retargeting can be performed online. Mathematically, we have

$$\left|\frac{\partial s_t}{\partial t}\right| \leq v^z_{\max}, \left\|(\frac{\partial x_t}{\partial t}, \frac{\partial y_t}{\partial t})\right\| \leq v_{\max}, \left|\frac{\partial^2 s_t}{\partial t^2}\right| \leq a^z_{\max}, \left\|(\frac{\partial^2 x_t}{\partial t^2}, \frac{\partial^2 y_t}{\partial t^2})\right\| \leq a_{\max} \tag{5}$$

where $v^z_{\max}$, $v_{\max}$, $a^z_{\max}$ and $a_{\max}$ are the maximum zooming and motion speed, and the maximum zooming and motion acceleration during cropping and scaling, respectively. Such first and second order constraints ensure that the view movement of the retargeted video is small, and ensure that there is no abrupt change of motion or zooming directions. They are both essential to the aesthetics of the retargeted video. Additional constraints may be derived from rules suggested by professional videographers [7]. It is our future work to incorporate these professional videography rules.

3 Detecting and tracking salient regions

We develop a two stage coarse-to-fine strategy for detecting and tracking salient regions, which is composed of an efficient exhaustive coarse search, and a gradient-based fine search as well. Since this two stage search process is performed at each time instant, to simplify the notation and without sacrificing clarity, we shall leave out the subscript t for some equations in the rest of this section.

Both search processes are facilitated by integral images, we employ the following notations for the integral image [14] of the saliency image $\mathcal{S}(x, y)$ and its partial derivatives, i.e., $\mathcal{T}(x, y) = \int_0^x \int_0^y \mathcal{S}(x, y)dxdy$, $\mathcal{T}_x(x, y) = \frac{\partial \mathcal{T}}{\partial x} = \int_0^y \mathcal{S}(x, y)dy$, and $\mathcal{T}_y(x, y) = \frac{\partial \mathcal{T}}{\partial y} = \int_0^x \mathcal{S}(x, y)dx$. All these integral images can be calculated very efficiently by accessing each image pixel only once. We further denote $\hat{x}(x, s) = x + sW_r$, and $\hat{y}(y, s) = y + sH_r$. Using $\mathcal{T}(x, y)$, the cropping information loss can be calculated in constant time, i.e., $\mathbf{L}_c(s, x, y) = 1 - (\mathcal{T}(\hat{x}, \hat{y}) + \mathcal{T}(x, y)) - (\mathcal{T}(\hat{x}, y) + \mathcal{T}(x, \hat{y}))$.

The calculation of the resizing information loss can also be speeded up greatly using integral images. We introduce the squared difference image $D_s(x, y)$ for scaling by s as $D_s(x, y) = (I(x, y) - I_s(x, y))^2$. We then also define the integral images of $D_s(x, y)$ and its partial derivatives, which are denoted as $\mathcal{D}^s(x, y)$, $\mathcal{D}^s_x(x, y)$, and $\mathcal{D}^s_y(x, y)$. We immediately have $\mathbf{L}_r(s, x, y) = (\mathcal{D}^s(\hat{x}, \hat{y}) + \mathcal{D}^s(x, y)) - (\mathcal{D}^s(\hat{x}, y) + \mathcal{D}^s(x, \hat{y}))$. In run time, we keep a pyramid of the integral images of $D^s(x, y)$ for multiple s. Since both $\mathbf{L}_c$ and $\mathbf{L}_r$ can be calculated in constant time, we are able to afford the computation of an exhaustive coarse search over the solution space for the optimal cropping window.

Once we have coarsely determined the location of a cropping window $\mathcal{W}$, we further exploit a gradient-based search to refine the optimal cropping window. By simple chain rules, it is easy to figure out that $\frac{\partial \mathbf{L}}{\partial a} = \mathcal{T}_a(\hat{x}, y) + \mathcal{T}_a(x, \hat{y}) - \mathcal{T}_a(x, y) - \mathcal{T}_a(\hat{x}, \hat{y}) + \lambda[\mathcal{D}^s_a(\hat{x}, y) + \mathcal{D}^s_a(x, \hat{y}) - \mathcal{D}^s_a(x, y) - \mathcal{D}^s_a(\hat{x}, \hat{y})]$, for $a = x$ or $a = y$, and $\frac{\partial \mathbf{L}}{\partial s} = A(x, y, s)W_r + B(x, y, s)H_r + \lambda \frac{\partial \mathbf{L}_r}{\partial s}$, where $A(x, y, s) = \mathcal{T}_x(\hat{x}, y) - \mathcal{T}_x(\hat{x}, \hat{y})$, $B(x, y, s) = \mathcal{T}_y(x, \hat{y}) - \mathcal{T}_y(\hat{x}, \hat{y})$, $\frac{\partial \mathbf{L}_r(x,y,s)}{\partial s} = \frac{\mathbf{L}_r(x,y,s+\triangle s) - \mathbf{L}_r(x,y,s-\triangle s)}{2\triangle s}$ is evaluated numerically. Then we perform a gradient descent step with backtracking line search to refine the optimal cropping window. Note that the gradient descent step is also very efficient because all derivatives can be calculated very efficiently using integral images and its partial derivatives. This two-step coarse-to-fine search ensures us to obtain the optimal cropping window very efficiently.

The feasible solutions $\Omega_t = \{[x^{min}_t,\ x^{max}_t],\ [y^{min}_t, y^{max}_t],\ [s^{min}_t,\ s^{max}_t]\}$ are derived from Eqs. 5 and strictly reenforced in tracking. Denote $\mathcal{W}^*_{t-1} = (x^*_{t-1}, y^*_{t-1}, s^*_{t-1})$ be the optimal cropping at the time instant $t - 1$, and let the optimal cropping window after these two stage search process at time instant t be $\hat{\mathcal{W}}_t$, we perform an exponential

moving average scheme to further smooth the parameters of the cropping window, i.e., $\mathcal{W}_t^* = \alpha\hat{\mathcal{W}}_t + (1-\alpha)\mathcal{W}_{t-1}^*$. We use $\alpha = 0.7 \sim 0.95$ in the experiments. It in general produces visually smooth and pleasant retargeted video, as shown in our experiments.

4 Scale-space spatiotemporal saliency

We propose several extensions of the spectrum residue method for saliency detection proposed by Hou and Liu [13]. We refer the readers to [13] for the details of their algorithm. Fig. 4(a) presents one result of saliency detection using the spectrum residue method proposed in [13]. On one hand we extend the spectrum residue method temporally, and on the other hand, we extend it in scale-space. The justification of our temporal extension may largely be based on the statistics of optical flows in natural images revealed by Roth and Black [15], which shares some common characteristics with the natural image statistics. It is also revealed by Hou and Liu [13] that when applying the spectrum residue method to different scales of the same image, different salient objects of different scales will pop out. Since for retargeting, we would want to retain salient object across different scales, we aggregate the saliency results from multiple scales together to achieve that.

Moreover, we also found that it is the phase spectrum [16] which indeed plays the key role for saliency detection. In other words, if we replace the magnitude spectrum residue with constant 1, the resulted saliency map is almost the same as that calculated from the spectrum residue method. We call such a modified method to be the *phase spectrum* method for saliency detection. The difference of the resultant saliency maps is almost negligible but it saves significant computation to avoid calculating the magnitude spectrum residue, as we clearly demonstrate in Fig. 4. Fig. 4(a) is the saliency map obtained from the spectrum residue and Fig. 4(b) is the saliency map produced from the phase spectrum only. Note the source image from which these two saliency maps are generated is presented as the top image in Fig. 3(a). The difference is indeed tiny. This is a common phenomenon that has been verified constantly in our experiments.

More formally, let $\mathcal{V}_t^n(i,j,k) = \{I_{t-n+1}(i,j), I_{t-n+2}(i,j), \ldots, I_t(i,j)\}$ be a set of n consecutive image frames and k indexes the image. Denote $\mathbf{f} = (f_1, f_2, f_3)$ as the frequencies in the fourier domain, where (f_1, f_2) represents spatial frequency and f_3 represents temporal frequency. The following steps are performed to obtain the spatiotemporal saliency map for $\mathcal{V}_t^n$:

1. Let $\Theta(\mathbf{f}) = Pha(\mathfrak{F}[\mathcal{V}_t^n])$ be the phase spectrum of the 3D FFT of $\mathcal{V}_t^n$.
2. Perform the inverse FFT and smoothing, i.e., $\mathbf{S}_t(i,j,k) = g(i,j) * \mathfrak{F}^{-1}\left[\exp\{j\Theta(\mathbf{f})\}\right]^2$. The smoothing kernel $g(i,j)$ is applied only spatially, since the temporal information will be aggregated.
3. Combine $\mathbf{S}(i,j,k)$ to be one single map, i.e., $\mathcal{S}_t(i,j) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{S}_t(i,j,k)$

The above steps present how to compute the spatiotemporal saliency map at a single scale. We aggregate the visual saliency information calculated from multiple scales together, this leads to the *scale-space spatiotemporal saliency*. More formally, let $\mathcal{V}_t^n(s)$ be the down-sampled version of $\mathcal{V}_t^n$ by a factor of s, i.e., each image in $\mathcal{V}_t^n$ is down-sampled by a factor of s in $\mathcal{V}_t^n(s)$. Denote $\mathcal{S}_t^s(i,j)$ as the spatiotemporal saliency image

calculated from $\mathcal{V}_t^n(s)$ based on the algorithm presented above. We finally aggregate the saliency map across different scales together, i.e.,$\mathcal{S}_t(i,j) = \frac{1}{n_s}\sum_s S_t^s(i,j)$, where n_s is the total number of levels of the pyramid. Fig. 3 presents the results of using the proposed approach to scale-space spatiotemporal saliency detection. The current image frame is the top one showing in Fig 3(a). We highlight the differences between the scale-space spatiotemporal saliency image (Fig. 3(c)) and the saliency maps (Fig. 4(a) and (b)) produced by the spectrum residue method [13] and the phase spectrum method, using color rectangles.

The proposed method successfully identified the right arm (*the red rectangle*) of the singer as a salient region, while the saliency map in Fig. 4(a) and (b) failed to achieve that. The difference comes from the scale-space spatiotemporal integration (the arm is moving) of saliency information. Moreover, in the original image, the gray level of the string in the blue rectangle is very close to the background. It is very difficult to detect its saliency based only on one image (Fig. 4 (b)). Since the string is moving, the proposed method still successfully identified it as a salient region (Fig. 3 (c)).

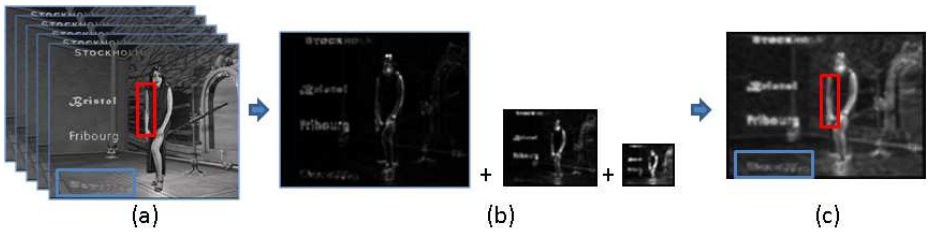

Fig. 3. Scale-space spatiotemporal saliency detection.

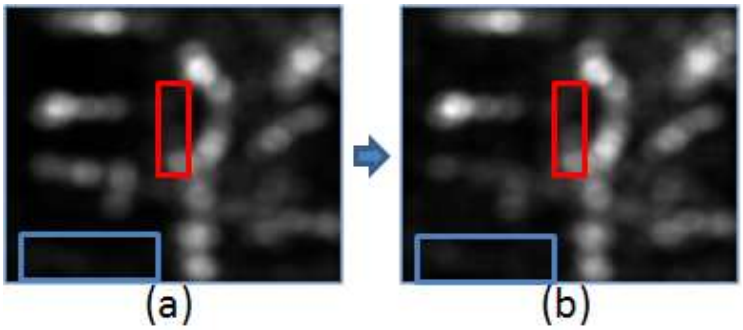

Fig. 4. Saliency detection using (a) spectrum residue [13], and (b) phase spectrum. The source image is shown in Fig. 3(a).

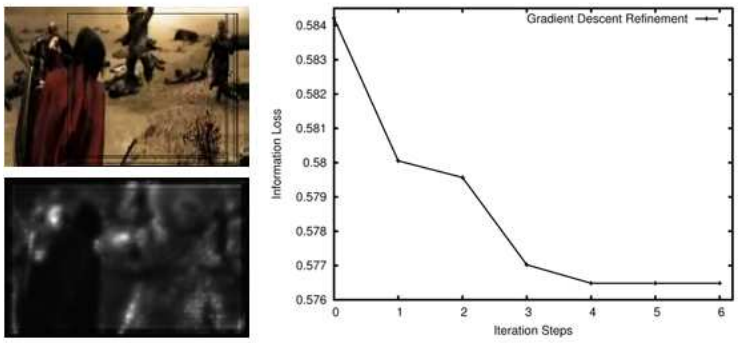

Fig. 5. Left column: the source image and its saliency map. Right column: the progress of the gradient search.

Fig. 6. Retargeting from 368×240 to 132×120 for movie video "300". The first four columns present the saliency tracking results and the corresponding saliency map. The fifth column shows our retargeting results. The sixth column shows the results by directly scaling.

5 Experiments

The proposed approach is tested on different videos for various retargeting purpose, including both standard MPEG-4 testing videos and a variety of videos downloaded from the Internet. All experiments are running with $\lambda = 0.3$ in Eq.2, which is empirically determined to achieve a good tradeoff. Furthermore, $n = 5$ video frames and an $n_s = 3$ level pyramid are used to build the scale-space spatiotemporal saliency map. We recommend the readers to look into the supplemental video for more details of our experimental results.

5.1 Spatiotemporal saliency tracking

To better understand the proposed approach to scale-space spatiotemporal saliency detection and tracking, we show a retargeting example on a video sequence from the battle scene of the movie "300". The video sequence has 1695 frames in total, we present some sample results in Fig. 6. As we can clearly see, the proposed saliency detection and tracking algorithms successfully locked onto the most salient regions. The fifth column of Fig. 6 presents our retargeting results. For comparison, the sixth column of Fig. 6 shows the results of directly resizing the original image frame to the target size. It is clear that in our retargeting results, the objects look not only larger but also keep their original aspect ratios even though the image aspect ratio changed from 1.53 to 1.1. To demonstrate the effectiveness of the gradient-based refinement step, we present the intermediate results of the gradient search at frame $\#490$ in in Fig. 5.

5.2 Content-aware v.s. content independent resizing cost

One fundamental difference between our approach and Liu and Gleicher [1] is that our resizing cost (Eq. 4) is dependent on the content of the cropped image. In contrast Liu and Gleicher only adopt an naive cubic loss $(s - 1.0)^3$ to penalize large scaling. To better understand the difference, we implemented a different retargeting system by replacing Eq. 4 with the naive cubic loss. The other steps remain the same. Therefore the differences in results are solely decided by the two different resizing costs. We call them *content aware* scheme and *content blind* scheme, respectively.

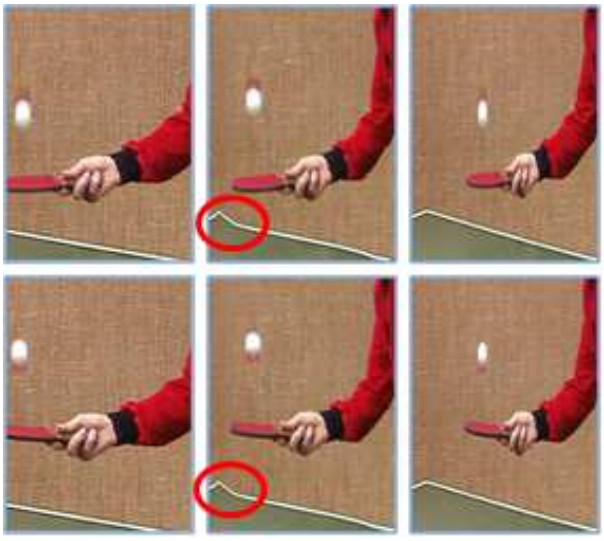

Fig. 7. Retargeting MPEG-4 standard test sequence "tennis". From left to right: **first column**–our approach, **second column**– Wolf et. al[2]'s method (by courtesy), **third column**– direct scaling.

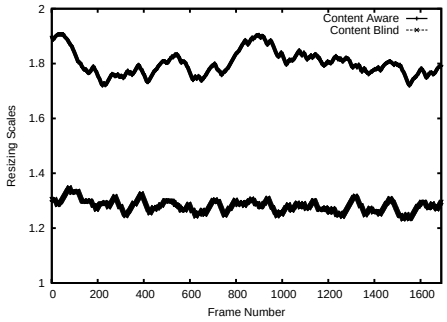

Fig. 8. The scaling factors associated with each video frame of the retargeting video "300". **Top**: content aware; **Bottom**: content blind.

Fig. 9. Retargeting MPEG-4 standard test sequence "Akiy" to be half of its original size: (a) direct scaling; (b) proposed approach; (c) Wolf et. al [2] (by courtesy).

We analyze the behaviors of the two methods based on the retargeting results of "300" video. Both cost values are normalized to be between 0 and 1 for fair comparison. For the content blind scheme, the λ is empirically determined on this video to be 0.2 for the best retargeting result. All other parameters are the same for the two methods. The curves in the upper and lower part of Fig. 8 present the scaling parameters from content aware resizing, and content blind resizing across the video, respectively.

It is clear that the content blind loss strongly favors small scaling. This bias may be very problematic because of the potentially large cropping information loss. In contrast, the content aware resizing does not have such a bias and also shows much larger dynamic range. This indicates that it is more responsive to capture the video content change. To achieve good results, we find that for the content blind scheme, the λ needs to be carefully tuned for each video, and its variance is large across different videos. In contrast, for the content aware scheme, a constant $\lambda = 0.3$ usually works well.

5.3 Video re-targeting results

We tested the proposed approach in a wide variety of long range video sequences for different retargeting tasks. We mainly show the retargeting results from the source video to 128×160 displays (Motorola-$T7xx$, NEC-$5x5$, SonyEricsson-$T610$,$T620$, SumSung-$VI660$) or 128×128(SumSung-$E175$, SonyEriccson-$Z200$), since these are the two widely adopted resolutions for mobile phones.

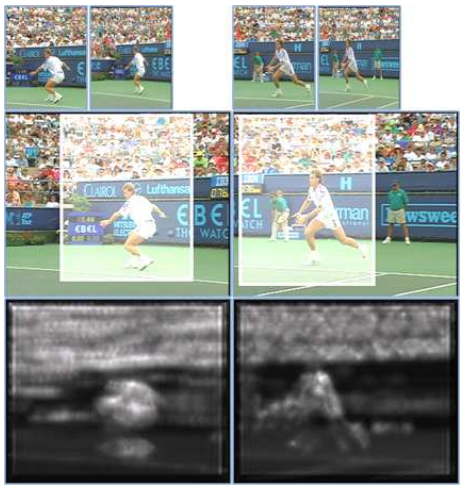

Fig. 10. Retargeting to 128×160.

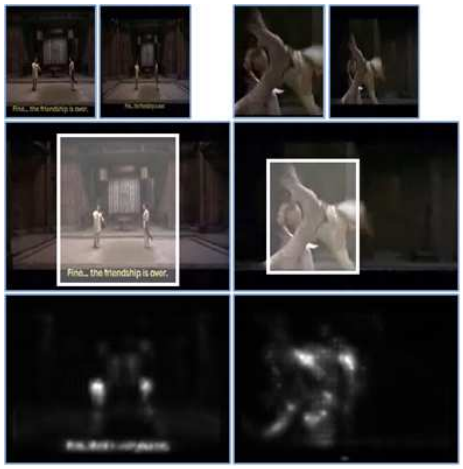

Fig. 11. Retargeting to 128×160.

The first retargeting result we present is performed on the standard MPEG-4 test video sequence"tennis". We re-target the source video to 176×240. The retargeted results from our approach on frame $\#10$ and $\#15$ are shown in the first column of Fig. 7. For comparison, we also present the retargeting results from Wolf et. al[2][1], and the results by direct scaling, in the second and third columns of Fig. 7, respectively. Due to the nonlinear warping of image pixels in Wolf et. al's method [2], visually disturbing distortion appears, as highlighted by the red circles in Fig. 7. In Fig. 9, we further compare our results with Wolf et. al [2] on the standard MPEG-4 testing video "Akiy". The task is to re-target the original video down to half of its original width and height. As we can clearly observe, the retargeted result from Wolf et. al [2] (Fig. 9 (c)) induces heavy nonlinear distortion, which makes the head size of the person in the video to be unnaturally big compared to her body size. In contrast, the result from the proposed approach keeps the original relative size and distortion free. Moreover, compared with the result from the direct scaling method in Fig. 9 (a), our result shows more details of the broadcaster's face when presented in a small display.

Fig. 10, Fig. 11, Fig. 12 and Fig. 13 present the video retargeting results on standard MPEG-4 testing video "stef", the best fighting scene of "Crouching Tiger" (2329 frames), a "Tom and Jerry" video (5692 frames), and a football video (517 frames). In all these figures, the first and third image in the first row presents the retargeting results from our approach, while the second and fourth images in the first row presents the results from direct scaling. The second and third row show the saliency tracking results, and the corresponding scale-space spatiotemporal saliency map, respectively. Compared with the direct scaling method, our retargeting results show significant better visual quality. In Fig. 11, when performing retargeting we purposely include the

[1] We thank Prof. Lior Wolf and Moshe Guttmann for their result figures.

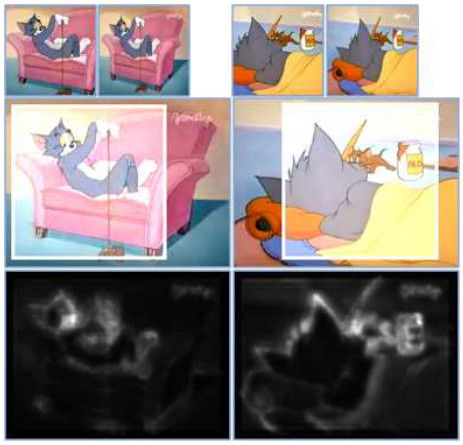

Fig. 12. Retargeting to 128×128.

Fig. 13. Retargeting to 128×128.

padding black bars in the original video to demonstrate the effectiveness of our saliency detection method. Notice how the caption text has been detected as salient region. These results demonstrate the advantages of the proposed approach. We strongly recommend the readers to watch our supplemental video for detailed results.

5.4 Effects of camera motion

Camera motion of the source video poses special challenges for video retargeting, especially for saliency detection. Previous approach for saliency detection, which combines motion cues with other visual cues such as image gradient and face detection results in an ad hoc manner, all suffer from this issue. Detecting and compensating camera motion is an effective way of resolving this issue. For example, Liu and Gleicher [1] perform motion segmentation to estimate the foreground motion more reliably. Our experiments show that the computational model introduced in 4 for scale-space spatiotemporal saliency detection is quite robust to camera motion. Fig. 14 presents the scale-space spatiotemporal saliency maps obtained on sample frames of a football video downloaded from YouTube (the source video is the same as the retargeting results we showed in Fig. 13). Both the original frame and the saliency map are presented. As we can observe, although the camera motion is very large, the foreground players still pop out as the most salient regions. The background advertisement board and the audiences are also somewhat salient because they are spatially salient. They are blurred a bit due to the camera motion, though. However, in our experiments, we found that such blurriness does not cause much issues, as shown in our retargeting results, e.g., on the "football" and "Tom& Jerry" sequence. While we agree that compensating camera motion may further improve the results, we trade it for computation efficiency and leave it to be our future work.

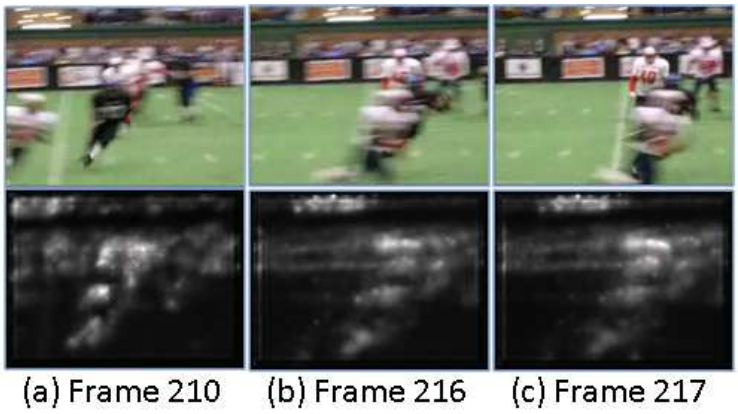

Fig. 14. The spatiotemporal saliency maps under heavy camera motions.

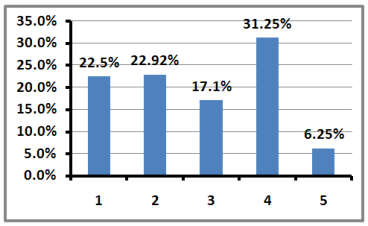

Fig. 15. The distribution of all the scores given by 30 users on 8 video clips. A score of 1 (5) is strongly positive (negative) about our approach.

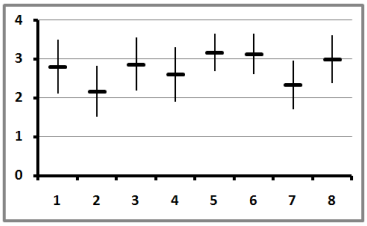

Fig. 16. The score of each individual video clip. The horizontal bars and vertical lines show the average scores and the standard deviations.

5.5 User study

We also performed a user study to evaluate the results. Without revealing to the users which results are from which methods, we ask the participants to look side-by-side the retargeting results on 8 video clips from the proposed approach, and those from the direct scaling (please refer to the supplemental video, in which video clips are shown in the same order as in our user study.). The users then mark in 5 scales regarding their preferences of the results, with 1 being preferring much more of the proposed approach, 3 being neutral, and 5 being preferring much more of the direct scaling approach. So the smaller the score, the more preference over the results from the proposed approach. There are 30 users with various background who participated in our user study.

We first present the distribution of all the scores over the 8 clips from all the 30 users in Fig. 15. Over all the scores, 22.5% strongly prefer and 22.92% moderately prefer the retargeted video from our approach, which add up to 45.42%. While 17.08% vote that our results and the results from direct scaling are almost the same. In contrast, there are also 31.25% moderately prefer and only 6.25% strongly prefer the direct scaling

results, i.e., 37.5% in total. This shows that 62.50% of the time, the users would feel that the results from the proposed approach are better or not worse than those from direct scaling. We also present the mean scores and standard deviations of each test video clip in Fig. 16. In total five clips got average scores lower than 3, two clips got average scores slightly higher than 3, and the last one got an average score of 3. This also manifests that users generally prefer the retargeting results from the proposed approach.

6 Conclusion and future work

We proposed a novel approach to distortion-free video retargeting by scale-space spatiotemporal saliency tracking. Extensive evaluation on a variety of real world videos demonstrate the good performance of our approach. Our user study also provide strong evidences that users prefer the retargeting results from the proposed approach. Future works may include further investigating possible means of integrating more professional videography rules into the proposed approach.

References

1. Liu, F., Gleicher, M.: Video retargeting: automating pan and scan. In: Proc. ACM international conference on Multimedia, ACM (2006) 241–250
2. Wolf, L., Guttmann, M., Cohen-Or, D.: Non-homogeneous content-driven video-retargeting. In: Proceedings IEEE International Conference on Computer Vision. (2007)
3. Setlur, V., Takagi, S., Raskar, R., Gleicher, M., Gooch, B.: Automatic image retargeting. In: Proc. International Conference on Mobile and Ubiquitous Multimedia. (2005)
4. Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. ACM Multimedia Systems Journal **9** (2003) 353–364
5. Luis Herranz, J.M.M.: Adapting surneillance video to small displays via object-based cropping. In: Proc. International Workshop on Image Analysis for Multimedia Interactive Services. (2007) 72–75
6. Liu, H., Xie, X., Ma, W.Y., Zhang, H.J.: Automatic browsing of large pictures on mobile devices. In: Proc. ACM international conference on Multimedia, ACM (2003)
7. Rui, Y., Gupta, A., Grudin, J., He, L.: Automating lecture capture and broadcast: technology and videography. ACM Multimedia Systems Journal **10** (2004) 3–15
8. Kang, H.W., Matsushita, Y., Tang, X., Chen, X.Q.: Space-time video montage. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Volume 2. (2006) 1331–1338
9. Gal, R., Sorkine, O., Cohen-Or, D.: Feature-aware texturing. In: Proceedings of Eurographics Symposium on Rendering. (2006) 297–303
10. He, L., Cohen, M.F., Salesin, D.: The virtual cinematographer: A paradigm for automatic real-time camera control and directing. In: Proc. Annual Conference on Computer Graphics (SIGGRAPH), ACM (1996) 217–224
11. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. ACM Transaction on Graphics, Proc. of SIGGRAPH'2007 **26** (2007) 10
12. Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. ACM Transaction on Graphics, Proc. of SIGGRAPH'2008 (2008)
13. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (2007)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Volume 1. (2001) 511–518
15. Roth, S., Black, M.J.: On the spatial statistics of optical flow. In: Proc. IEEE International Conference on Computer Vision. Volume 1. (2005) 42–49
16. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Volume 2. (2008) 1–8