An Enhanced Seam Carving Approach for Video Retargeting

Tzu-Hua Chao, Jin-Jang Leou, and Han-Hui Hsiao
Department of Computer Science and Information Engineering
National Chung Cheng University, Chiayi 621, Taiwan
E-mail: {cth98m, jjleou, hhh95p}@cs.ccu.edu.tw Tel: +886-5-2720411 ext. 33105

***Abstract*—Video retargeting (resizing) is an important task for displaying videos on various display devices. In this study, an enhanced seam carving approach for video retargeting is proposed, in which a seam may be a non-8-connected one. Both the search window size and the temporal weight can be adaptively adjusted according to video contents (motion information). Additionally, to preserve temporal coherence, the appearance-based method is employed. The spatial and temporal costs of a pixel are linearly combined to compute the cumulative cost with an adaptive temporal weight. Finally, dynamic programming is used to determine the optimal non-8-connected seam (with the minimum cumulative cost) for carving out. Based on the experimental results obtained in this study, the performance of the proposed approach is better than those of two comparison approaches.**

I. INTRODUCTION

Video retargeting (resizing) is an important task for displaying videos on various display devices [1]. Existing video retargeting approaches are generally classified into four categories, namely, cropping, warping, seam carving, and their combinations. Many video retargeting approaches are addressed by extending some image retargeting approaches with some temporal conditions, in which video frames are processed independently, resulting in temporal artifacts.

For cropping approaches, Liu and Gleicher [1] proposed an automating pan and scan method, in which each frame is cropped and scaled to the new size by estimating the important information in videos. The cropping window may be moved during a shot to introduce virtual pans and cuts. Similarly, Deselaers et al. [2] optimized each cropping sequence over time to preserve temporal coherence. Other similar methods [3-5] used different techniques to determine the optimized trajectory for a cropping window in a video sequence.

For warping approaches, visually important regions will be uniformly scaled, while unimportant regions may be warped with arbitrary deformations [5]. Krähenbühl et al. [5] proposed a novel and integrated system for context-aware video retargeting. Wolf et al. [6] proposed a non-homogeneous content-aware video retargeting approach. Local saliency, motion detection, and face detection are extracted as important information. Then, temporal coherence is preserved by constraining transformed pixels between two consecutive frames. Wang et al. [7] built a video retargeting framework by incorporating some motion-aware constraints. Niu et al. [8] proposed an effective frame-based warping approach, in which a motion history map is used to propagate information about moving objects between video frames.

For seam carving approaches, Rubinstein et al. [9] proposed a novel energy function, which improves the visual quality of the retargeting images and videos. The new energy function is looking "forward" in time, i.e., removing seams that introduce the least amount of energy into the retargeting results. To achieve video retargeting, they apply graph cuts to find 2-D seams (surfaces) in a space-time volume. Grundmann et al. [10] proposed a frame-based seam carving approach, which relies on a novel appearance-based temporally coherent formulation. Kopf et al. [11] proposed a fast video seam carving approach for handheld mobile devices.

Additionally, Wang et al. [12] proposed a motion-based video retargeting approach with optimized crop-and-warp. Rubinstein et al. [13] proposed a multi-operator video retargeting approach, which combines seam carving, cropping, and scaling to produce final video results. A new image similarity measure is defined and dynamic programming is used to final an optimal path in the resizing space.

The paper is organized as follows. The proposed video retargeting approach is described in Section II. Experimental results are included in Section III, followed by concluding remarks.

II. PROPOSED APPROACH

The framework of the proposed approach is shown in Fig. 1. The spatial and temporal costs of each pixel in the current frame are computed by the spatial and temporal energy functions, respectively. The spatial cost reflects spatial visual artifacts, while the temporal cost represents temporal coherence. The spatial and temporal costs of a pixel can be linearly combined with different weights to compute the cumulative cost. Finally, using dynamic programming [9], the

+ This work was supported in part by National Science Council, Taiwan, Republic of China under Grants NSC 99-2221-E-194-032-MY3 and NSC 101-2221-E-194-031.

seam with the minimum cumulative cost is iteratively determined and carved out until the desired frame size is reached.

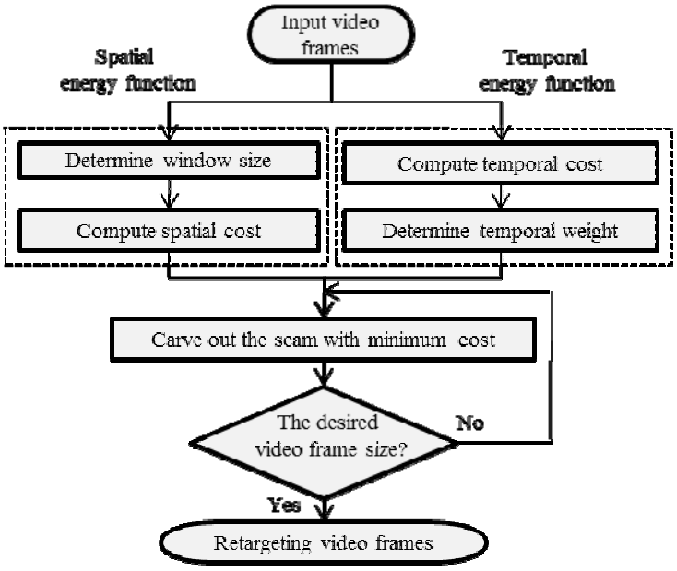


Fig. 1. The framework of the proposed approach.

A. Spatial Cost and Window Size Determination

In this study, the spatial energy function reflecting the importance of a pixel in a video frame [10] is employed, which may produce a non-8-connected seam. To determine an optimal non-8-connected seam, all pixels in the current row may be searched, which is computationally expensive. To reduce computational complexity, an adaptive search window is employed. The spatial cost S_c of a pixel contains two terms, S_h and S_v, which reflect the visual errors in the horizontal and vertical directions, respectively, if the pixel is removed, i.e., $S_c = S_h + S_v$.

S_h of a pixel is defined as the difference of the intensity gradients in the horizontal direction if the pixel is removed. As two illustrated examples shown in Fig. 2, the spatial cost $S_h(A)$ for removing a border pixel A (Fig. 2 (a)), i.e., the visual error in the horizontal direction, is $S_h(A)=\big||A-B|-|B-C|\big|$. Similarly, the spatial cost $S_h(B)$ for removing an inside pixel B (Fig. 2 (b)), i.e., the visual error in the horizontal direction, is $S_h(B)=|A-B|+|B-C|-|A-C|$.

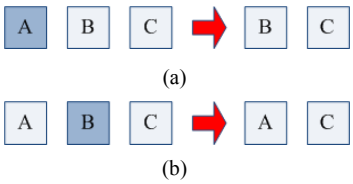


Fig. 2. Two illustrated examples for S_h: (a) removing a border pixel A; (b) removing an inside pixel B [10].

On the other hand, S_v is a transition cost between a pair of removing pixels in adjacent rows. An illustrated example for S_v is shown in Fig. 3. Suppose that pixel E will be removed. In Fig. 3 (a), pixel A is removed in the previous row and S_v of pixel E is $S_v(E,A)=\big||A-D|-|B-D|\big|+\big||B-E|-|B-D|\big|$. In Fig. 3 (b), pixel C is removed in the previous row, and S_v of pixel E is $S_v(E,C)=\big||C-F|-|B-F|\big|+\big||B-E|-|B-F|\big|$.

For a non-8-connected seam, the transition cost will accumulate transition costs. In Fig. 3 (c), pixel A is removed in the previous row, and the "cumulative" transition cost S_v of pixel F (for the case $A<F$) is $S_v(F,A)=\big||A-\mathrm{D}|-|B-D|\big|+\big||B-E|-|C-E|\big|+\big||C-F|-|C-E|\big|+\big||B-E|-|B-D|\big|$. Fig. 3 (d) shows the cumulating process for $S_v(F,A)$ in Fig. 3 (c).

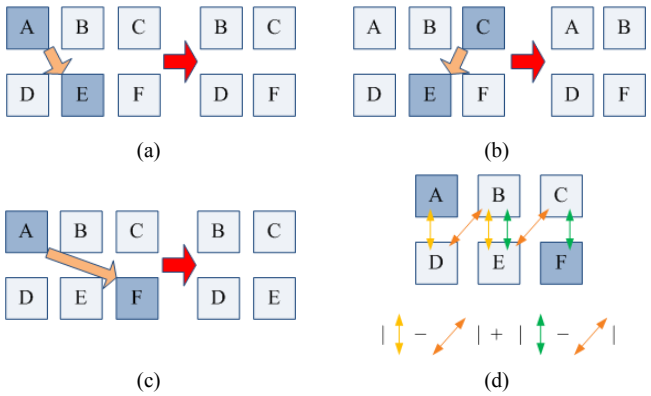
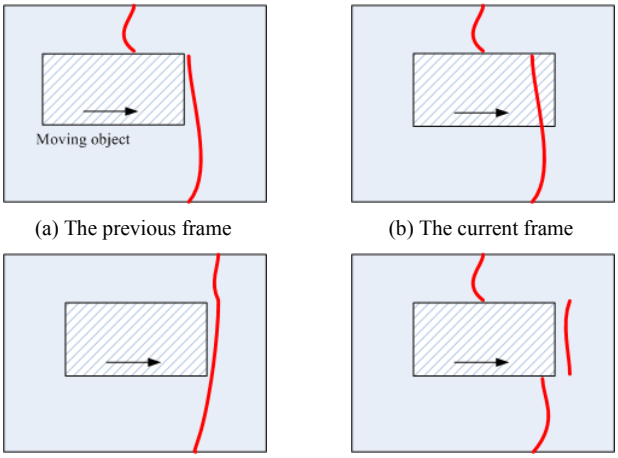


Fig. 3. Three illustrated examples for S_v: (a) pixel A in the previous row and pixel E in the current row are removed; (b) pixel C in the previous row and pixel E in the current row are removed; (c) pixel A in the previous row and pixel F in the current row are removed; (d) the cumulating process for $S_v(F,A)$ in Fig. 3(c) [10].

To determine an optimal non-8-connected seam, all pixels in the current row may be searched, which is computationally expensive. To reduce computational complexity, the search range in [10] is restricted to 15 pixels. Note that a removed seam may damage a moving object, if the search window size is too small. In Fig. 4 (a), the moving object may not be damaged in the previous frame when the moving object is moving from left to right. In Fig. 4 (b), the moving object would be damaged in the current frame since the search window size is too small to preserve temporal coherence. In Fig. 4 (c), the moving object would not be damaged in the current frame, but temporal coherence may not be preserved. In Fig. 4 (d), the moving object would not be damaged if search window size is large enough for preserving moving object's shape and temporal coherence.

The computational complexity for determining an optimal non-8-connected seam depends on the search window size. In this study, the search window size of the current frame will be adaptively adjusted according to video contents (motion information). Here, motion vectors (MVs) obtained by motion estimation between two consecutive frames are employed. For example, if the frame width will be reduced (resized), the horizontal motion vectors are considered only. $MV_i(x, y)$ denotes the horizontal motion vector of pixel (x, y) in frame i,

which will predict the horizontal motion of the pixel in frame $i+1$.



(a) The previous frame (b) The current frame

(c) The current frame (d) The current frame

Fig. 4. The moving object may be damaged by carving out the seam (in red): (a) the seam can be non-8-connected so that the moving object mat not be damaged; (b) the moving object would be damaged in the current frame because the search window size is too small; (c) the moving object would not be damaged in the current frame, but temporal coherence may not be preserved; (d) the moving object would not be damaged if the search window size is large enough for preserving moving object's shape and temporal coherence.

Suppose that a seam S_{i-1} is found previously in frame $i-1$, and the seam should be determined in the current frame. The motion information of the seam should be identified around in frame $i-1$. The identified range is $2\times\left\lfloor\frac{w}{32}\right\rfloor$ pixels, where w is the width of the current frame. $N_i(y)$, the number of motion pixels around the seam, is given as

$$N_i(y)=\sum_{t=S_{i-1}(y)-\left\lfloor\frac{w}{32}\right\rfloor}^{S_{i-1}(y)+\left\lfloor\frac{w}{32}\right\rfloor} C_{i-1}(t,S_{i-1}(y),y), \tag{1}$$

$$C_{i-1}(t,x,y)=\begin{cases}1, & \text{if } (MV_{i-1}(t,y)-MV_{i-1}(x,y))\neq 0,\\ 0, & \text{otherwise,}\end{cases} \tag{2}$$

where $S_{i-1}(y)$ denotes the x-axis of the seam in the height y of frame $i-1$. The motion ratio $r_i(y)$ is obtained by

$$r_i(y)=\frac{N_i(y)}{2\times\left\lfloor\frac{w}{32}\right\rfloor+1}. \tag{3}$$

After scanning each row, the maximum and minimum motion ratios r_i^{max} and r_i^{min} are obtained. The motion information of the seam is $r_i^{diff}=r_i^{max}-r_i^{min}$. Finally, the search window size WS_i in frame i is determined as

$$WS_i=\max\left(WS_{i-1}+\left\lfloor\frac{r_i^{diff}-r_{i-1}^{diff}}{0.1}\right\rfloor, WS_{i-1}\right). \tag{4}$$

Based on Eq. (4), the search window size may grow to contain all pixels of a row. To prevent this situation, after the seam is found in frame i, WS_i is refined as

$$WS_i=\max(S_i(y)-S_i(y+1)),\quad y\in 1,2,\ldots,h-1, \tag{5}$$

where h is the height of the current frame.

B. Temporal Cost

To preserve temporal coherence, an enhanced version of the method developed in [10] is employed, which measures the visual difference between the optimal temporally coherent frame and a resulting frame by carving out a seam. Note that the seams in consecutive frames may be not geometrically smooth.

Suppose that a seam S_{i-1} is found previously in frame $i-1$. We will carve out a seam from frame i so that the resulting frame would be visually close to the optimal temporally coherent frame I_i^c, which is obtained by carving out the previous seam S_{i-1} from frame i. The sum of squared differences of the two involved rows is employed to define temporal coherence $T_c(x,y)$ as

$$\begin{aligned}T_c(x,y)=&\sum_{k=1}^{x-1}(I_i(k,y)-I_i^c(k,y))^2\\&+\sum_{k=x+1}^{w-1}(I_i(k,y)-I_i^c(k-1,y))^2,\end{aligned} \tag{6}$$

where $I_i(x,y)$ is intensity of pixel (x,y) of frame i and w is the width of frame i.

The seams carved out in the previous frame would influence the seams in the current frame. The retargeting frame may be optimal for considering both the spatial and temporal costs, but it may be sub-optimal for the spatial cost. In [10], the temporal weight is 0.2 when video contents are highly dynamic, and the temporal weight is 1 for most video sequences. In this study, the temporal weight should be adaptively adjusted according to video contents (motion information).

Suppose that the frame width would be reduced, horizontal motion vectors are considered only. $MV_i(x,y)$ denotes the horizontal motion vector of pixel (x,y) in frame i, which can predict the horizontal motion of the pixel in frame $i+1$. The motion information of the seam S_{i-1} should be also identified around frame $i-1$. The identified range is $\left\lfloor\frac{w}{32}\right\rfloor$ pixels, where w is the width of the current frame. Then, $N_i^L(y)$ and $N_i^R(y)$ in the height y in frame i are defined, respectively, as

$$N_i^L(y)=\sum_{t=S_{i-1}(y)-\left\lfloor\frac{w}{32}\right\rfloor}^{S_{i-1}(y)-1} C_{i-1}^L(t,S_{i-1}(y),y), \tag{7}$$

$$C_{i-1}^L(t,x,y)=\begin{cases}1, & \text{if } (MV_{i-1}(t,y)-MV_{i-1}(x,y))>0,\\ 0, & \text{otherwise,}\end{cases} \tag{8}$$

$$N_i^R(y)=\sum_{t=S_{i-1}(y)+1}^{S_{i-1}(y)+\left\lfloor\frac{w}{32}\right\rfloor} C_{i-1}^R(t,S_{i-1}(y),y), \tag{9}$$

$$C_{i-1}^R(t,x,y)=\begin{cases}1, & \text{if } (MV_{i-1}(t,y)-MV_{i-1}(x,y))<0,\\ 0, & \text{otherwise,}\end{cases} \tag{10}$$

where $S_{i-1}(y)$ denotes the x-axis of the seam in the height y of frame $i-1$. $N_i^L(y)$ and $N_i^R(y)$ represent the numbers of pixels such that these pixels are moving closely to the seam in frame $i-1$ from left and right sides, respectively. Then, $N_i^L(y)$ and $N_i^R(y)$ divided by $\left\lfloor\frac{w}{32}\right\rfloor$ are the motion ratios $r_i^L(y)$ and $r_i^R(y)$, respectively. Finally, the temporal weight α is obtained by

$$\alpha=\begin{cases}0.05, & \text{if } r_i^L(y)>0.8 \text{ and } r_i^R(y)>0.8,\ y\in 1,2,\ldots,h,\\ 0.2, & \text{else if } r_i^{max}>0.8,\\ 1, & \text{otherwise,}\end{cases} \tag{11}$$

where r_i^{max} is the maximum motion ratio among the pixels in each row.

C. *Video Retargeting by Enhanced Seam Carving*

After the spatial and temporal costs of a pixel are determined, the two costs S_c and T_c are linearly combined to compute the cumulative cost M with an adaptive temporal weight α. Similar to [9], dynamic programming is used to determine the optimal non-8-connected seam with the minimum cumulative cost of each frame. Then, the video frame size will be reduced by iteratively carving out a seam with the minimum cumulative cost until the desired video frame size is reached.

III. EXPERIMENTAL RESULTS

In this study, the proposed video retargeting approach is implemented on an Intel Core 2 Duo E8400 3.0 GHz PC with 4GB main memory using Matlab of version 2010a software develop tool. Five test video sequences are employed. They are "persons" (320×240), "baseball" (384×216), "coach" (352×288), "basketball" (232×176), and "highway" (448×193). To evaluate the effectiveness of the proposed approach, linearly scaling and Grundmann et al.'s approach [10] are implemented in this study. On the other hand, the Rubinstein et al.'s retargeting results [9] are also compared for the video sequence "highway".

In Fig. 5. the original video frames of size 320×240 (without camera motions) are horizontally downsized to 200×240. The two persons in Fig. 5 (c) are distorted because the seams cannot jump to other locations due to fixed temporal weights. In Fig. 6, the original video frames of size 232×176 are horizontally downsized to 160×176. Because the video contents are highly dynamic, in Fig. 6 (c), the video contents are distorted due to camera motions and object motions. In Fig. 7, the original video frames of size 448×193 are horizontally downsized to 268×193. The video contents in Fig. 7 (b) are distorted and the objects are damaged, due to the seams in adjacent frames must be 8-connected. The retargeting results by Grundmann et al.'s approach [10] are similar to those of the proposed approach, while the proposed approach can preserve the details.

On the other hand, the computation times for 5 video sequences of Grundmann et al.'s approach [10] and the proposed approach are listed in Table 1. Grundmann et al.'s approach used the fixed search window size, which may need additional computation time. The proposed approach can adaptively adjust search window sizes according to video contents (motion information) so that the computational times can be reduced.

IV. CONCLUDING REMARKS

In this study, an enhanced seam carving approach for video retargeting is proposed, in which a seam may be a non-8-connected one. Both the search window size and the temporal weight can be adaptively adjusted according to video contents (motion information). To preserve temporal coherence, the appearance-based method is employed. The spatial and temporal costs of a pixel are linearly combined to compute the cumulative cost with an adaptive temporal weight. Finally, dynamic programming is used to determine the optimal non-8-connected seam (with the minimum cumulative cost) for carving out. Based on the experimental results obtained in this study, the performance of the proposed approach is better than those of two comparison approaches.

(a) (b) (c) (d)

Fig. 5. Experimental results of the video sequence "persons" (frames 8, 20, and 38): (a) the original video frames; (b)-(d) the video retargeting results by linear scaling, Grundmann et al.'s approach [10], and the proposed approach, respectively.

(a) (b) (c) (d)

Fig. 6. Experimental results of the video sequence "basketball" (frames 41, 102, and 111): (a) the original video frames; (b)-(d) the video retargeting results by linear scaling, Grundmann et al.'s approach [10], and the proposed approach, respectively.

(a) (b) (c) (d)

Fig. 7. Experimental results of the video sequence "highway" (frames 97, 102, and 111): (a) the original video frames; (b)-(d) the video retargeting results by Rubinstein et al.'s approach [9], Grundmann et al.'s approach [10], and the proposed approach, respectively.

TABLE I

THE COMPUTATION TIMES FOR 5 VIDEO SEQUENCES OF GRUNDMANN ET AL.'S APPROACH [10] AND THE PROPOSED APPROACH.

Video sequences	Number of frames	Width reduction	Computation times	
			Grundmann et al. [10]	Proposed
"persons"	40	37.5%	1539 (sec.)	1289 (sec.)
"baseball"	32	37.5%	1862 (sec.)	1590 (sec.)
"coach"	80	33.5%	5820 (sec.)	4311 (sec.)
"basketball"	228	31.0%	2441 (sec.)	2111 (sec.)
"highway"	194	40.1%	14593 (sec.)	13844 (sec.)

REFERENCES

[1] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," in *Proc. of ACM Int. Conf. on Multimedia*, 2006, pp. 241-250.

[2] T. Deselaers, P. Dreuw, and H. Ney, "Pan, zoom, scan – time-coherence, trained automatic video cropping," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 23-28.

[3] Y. Li, Y. Tian, J. Yang, L. Y. Duan, and W. Gao, "Video retargeting with multi-scale trajectory optimization," in *Proc. of ACM Int. Conf. on Multimedia Information Retrieval*, 2010, pp. 45-54.

[4] Z. Yuan, T. Lu, Y. Huang, D. Wu, and H. Yu, "Video retargeting: a visual-friendly dynamic programming approach," in *Proc. of IEEE Int. Conf. on Image Processing*, 2010, pp. 2857-2860.

[5] P. Krähenbühl and M. Lang, "A system for retargeting of streaming video," *ACM Trans. on Graphics*, vol. 28, no. 5, pp. 1-10, 2009.

[6] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous context-driven video-retargeting," in *Proc. of IEEE Int. Conf. on Computer Vision*, 2007, pp. 1-6.

[7] Y. S. Wang, H. Fu, O. Sorkine, T. Y. Lee, and H. P. Seidel, "Motion-aware temporal coherence for video resizing," *ACM Trans. on Graphics*, vol. 28, no. 5, pp. 1-10, 2009.

[8] Y. Niu, F. Liu, X. Li, and M. Gleicher, "Warp propagation for video resizing," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 537-544.

[9] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. on Graphics*, vol. 28, no. 3, pp. 1-9, 2009.

[10] M. Grundmann, V. Kwatra, M. Han, and I. Esssa, "Discontinuous seam-carving for video retargeting," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 569-576.

[11] S. Kopf, J. Kiess, H. Lemelson, and W. Effelsberg, "FSCAV – fast seam carving for size adaptation of videos," in *Proc. of ACM Int. Conf. on Multimedia*, 2009, pp. 321-330.

[12] Y. S. Wang, H. C. Lin, O. Sorkine, and T. Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," *ACM Trans. on Graphics*, vol. 29, no. 4, pp. 1-9, 2010.

[13] M. Rubinstein, A. Shamir, and S. Avidan, "Multi-operator media retargeting," *ACM Trans. on Graphics*, vol. 28, no. 3, pp. 1-11, 2009.