Content-Aware Video Seam Carving Based on Visual Cubes

Duan-Yu Chen* and Yi-Shiou Luo

Department of Electrical Engineering

Yuan-Ze University

Chung-Li 32003, Taiwan, ROC

dychen@saturn.yzu.edu.tw; s984629@mail.yzu.edu.tw

Received 18 April 2010; Revised 19 May 2010; Accepted 20 June 2010

Abstract. Seam carving for still images has attracted lots of attention in recent years. Approaches that can work well in this domain may not sufficiently robust enough to be applied to consecutive video frames due to the nature of visual dynamics in videos. Carving in consecutive frames with different criteria would usually result in discontinuity of visual perception. Therefore, how to preserve the visual continuity in video frames is the most critical issue in the field of video seam carving. In this paper, we propose a novel approach for modeling dynamic visual attention based on spatiotemporal analysis in order to detect the focus of interest automatically. The continuously varied co-sited blocks in a video cube are first detected and their variations are characterized as a bag of visual cubes, which are further employed to determine a proper extent of salient regions in video frames. Once the proper extent through video cubes is determined, the carving process then can be conducted to find the global optimum. Our experiment shows that the proposed content-aware video seam carving based on spatiotemporal bag of visual cubes can effectively generate resized videos while keeping their isotropic manipulation and the continuous dynamics of visual perception.

Keywords: Seam carving, visual cubes, spatiotemporal analysis

1 Introduction

The rapid growing of the diversity of devices that can browse multimedia has made video data accessible more and more easily. However, managing a huge amount of videos for displaying in distinct devices is a challenging task since for example the resolution of a video sequence has to be adjusted to fit the monitor size of the device. Furthermore, to provide better visual perception for users, the video size has to be dynamically adaptive to the size of a browsing window. Therefore, how to adapt the video content to a new display requirement while keeping its isotropic manipulation and the continuous dynamics of visual perception has become a critical research issue.

In recent years, seam carving for adapting still images has attracted lots of attention. Seam carving [1, 2] is a technique that is originally proposed to adjust the size of still images. The minimum energy in an image is computed iteratively to determine the positions where the pixels can be removed or inserted. However, to carve consecutive images in a video sequence, the temporal axis needs to be considered in order to preserve the continuity of visual perception. A good way to achieve this aim is to detect and protect the salient regions in 3D video volumes that users could be of interested during the carving process. However, what parts of a scene should be considered "salient"? According to a study conducted by cognitive psychologists [3] the human visual system picks salient features from a scene. Psychologists believe this process emphasizes the salient parts of a scene and, at the same time, disregards irrelevant information. To address this question, several visual saliency models have been proposed in the last decade [4-14]. Based on the type of attention pattern adopted, the models can be roughly categorized into two classes: bottom-up approaches, which extract image-based saliency cues; and top-down approaches, which extract task-dependent cues. Usually, extracting task-dependent cues requires a priori knowledge of the target(s). However, a priori knowledge of attended objects is usually difficult to obtain. Therefore, we focus on a bottom-up approach in this work.

In the related works of seam carving, Avidan and Shamir [1] define a vertical (horizontal) seam to be an 8-connected path in a raster scan order to find a pixel-wide path that is of the minimum gradient energy and then remove these pixels in the path. Thus, removing a vertical (horizontal) seam reduces the width (height) by one pixel. The approach in [1] can work well in finding the globally minimum energy seam while preserves salient regions in still images. For video seam carving, the additional temporal axis makes it more challenging than that

* Correspondence author

applied in still images. In [9], the resulting carved videos would show the discontinuity in consecutive frames, especially the broken salient objects.

Carving in consecutive frames using different criteria together, for example the spatial and temporal visual saliency, would usually result in visual discontinuity. Therefore, in this paper, we propose a novel approach for modelling dynamic visual attention based on spatiotemporal analysis in order to detect the focus of interest automatically. The continuously varied co-sited blocks in a video cube are first detected and their variations are characterized as a bag of visual cubes, which are further employed to determine a proper extent of salient regions in video frames. Once the proper extent through video cubes is determined, the carving process then can be conducted to find the global optimum. The concept of the proposed approach is shown in Fig.1. The seam surface for carving is determined by spatiotemporal analysis for protecting the salient visual cubes.

The remainder of the paper is organized as follows. In the next section, we describe the proposed visual saliency model and the method of video seam carving. Section 3 shows the experiment results. Then, in Section 4, we present our conclusions.

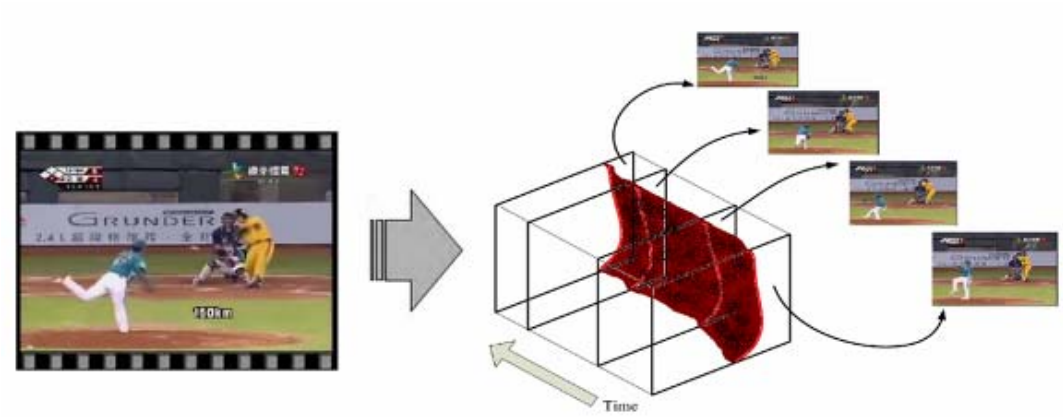


Fig. 1. The seam surface is determined by spatiotemporal analysis to preserve the visual continuity of salient visual cubes

2 Video Seam Carving Based on Bag of Visual Cubes

In this section, to preserve visual continuity during the process of video seam carving, we first compute the variations in 3D video volumes and then detect a bag of visual salient cubes. A visual salient map is accordingly constructed based on the positions of salient cubes. In order to address the priority of each cube, the importance of cubes is accumulated. Finally, the visual salient map is obtained and used for determining the carving surface. Section 2.1 introduces the method of detecting salient co-sited blocks in neighboring frames. Section 2.2 describes the features we computed in the salient blocks. Section 2.3 shows the method of determining a bag of salient cubes. Then, section 2.4 details the carving process based on the visual salient map.2.1.

2.1 Detecting Salient Regions in Neighboring Frames

Since the dynamic modeling of visual saliency is our concern, the variations of co-sited blocks in the neighboring frames are obtained by computing the difference between them. For reduce the effect of noises, each frame is first smoothed by a Gaussian kernel (a 3 × 3 or 5 × 5 binomial kernel) as defined by

$$\mathrm{D}(\mathrm{x},\ \mathrm{y}) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \qquad (\mathbf{1})$$

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution. Assume that we have a set of frames $f = \{1, 2, 3, ..., v\}$. The difference image I_{dg}, as demonstrated in Fig.2(c), between the original frames I^f and I^{f+1} can be obtained by 1-norm distance and then I_{dg} is binarized adaptively by Otsu's approach [15] to further filter out non-salient regions. In Fig.2(c), regions enclosing by a yellow circle are noises and the ones marked by a red bounding box are the targets.

To compute the gradient energy in the resulting binarized image, edges are first computed by using Sobel operator. The magnitude of the gradient

$$|G| = \sqrt{I_X^2 + I_Y^2}$$

is computed by using the horizontal and vertical gradients, I_X and I_Y, respectively. The orientation of the gradient is given by

$$\theta = \arctan\left[\frac{I_Y}{I_X}\right].$$

The process is shown in Fig.2. It can be observed that most salient regions can be detected in Fig. 2(e).

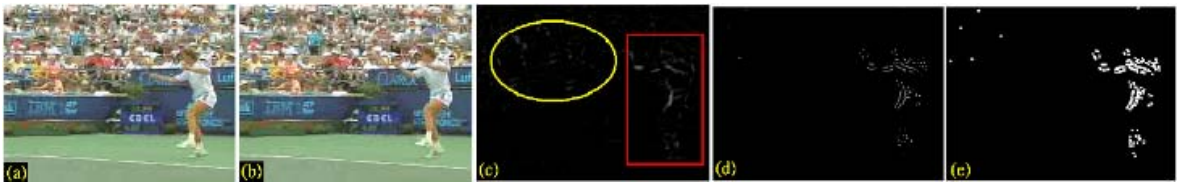


Fig. 2. Demonstration of the process of detecting salient regions (a) current frame; (b) next frame; (c) detected salient regions;(d) binarize (c); (e) edge detection

2.2 Detection of Spatiotemporal Salient Cubes

Detected salient regions have to be prioritized for further determination of carving surface. In 3D video volumes, gradients would change with the variation of moving targets and/or of the scene changes. Therefore, the difference of the gradient orientation between co-sited blocks is computed to prioritize detected salient regions through video frames.

We first compute the HOG [16] for each block. Fig.3. shows the concept that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions.

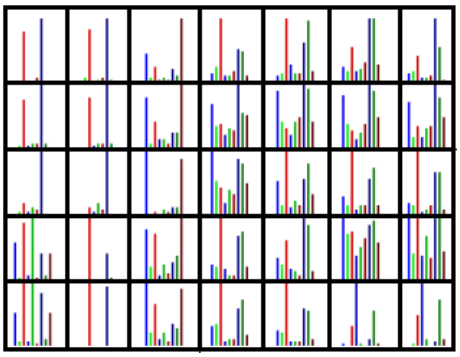


Fig.3. In practice this is implemented by dividing the image window into small spatial regions ("cells"), for each cell accumulating a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation

This can be done by accumulating a measure of local histogram "energy" over somewhat larger spatial regions ("blocks") and using the results to normalize all of the cells in the block. After the HOG of each frame is obtained, we can detect the spatiotemporal salient cubes. The HOG is used for comparing the co-sited blocks. Assume that each frame is equally divided into $M \times N$ blocks where each block is of size $m \times n$ and gradient energy of a block E. Considering the dynamics of co-sited blocks in neighbouring frames, we define the priority w of a block (M, N) by

$$w(M,N) = \sum_{i=0}^{m}\sum_{j=0}^{n} C(E_{(i,j)}^{f}, E_{(i,j)}^{f+1}), \quad (2)$$

$$C(E_{(i,j)}^{f}, E_{(i,j)}^{f+1}) = \begin{cases} 1 & different \\ 0 & otherwise \end{cases}, \tag{3}$$

where C is a response function to show the dynamics of the co-sited blocks, in which inconsistency between co-sited blocks would set its value to 1. Otherwise, C is set to 0 to indicate their orientation consistency. w is a priority function that can address the degree of visual saliency in 3D video volume. The approach is shown in Fig.4. A bag of visual salient cubes can then be obtained by detecting saliency between co-sited blocks when their corresponding value w is larger than a predefined threshold.

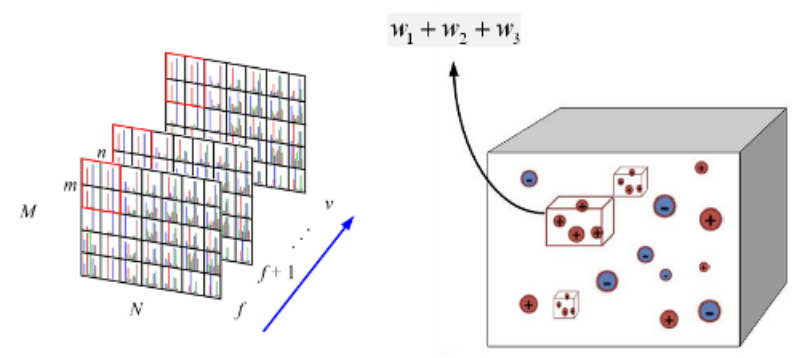


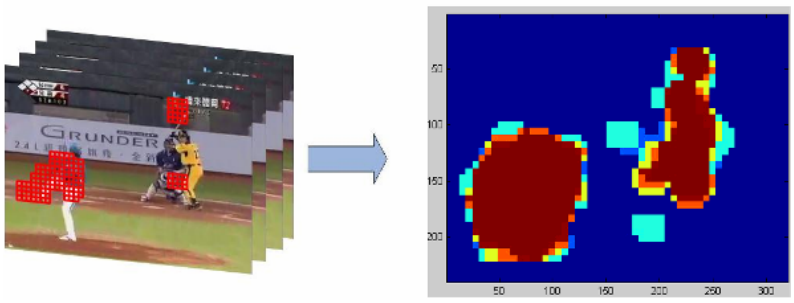
Fig. 4. Salient visual cubes with priority value w are obtained by computing the dynamics between co-sited blocks

2.3 Seam Carving Based on Bag of Visual Cubes

To determine the carving surface, it is necessary to first determine the extent of 3D salient regions. Using the detected cubes $\boldsymbol{W}$, the range of saliency SR is computed by

$$SR(M,N) = \sum_{i=1}^{f} W^{i}(M,N), \tag{4}$$

where $SR(M,N)$ is the degree of saliency of the block (M, N) and f denotes the frame number. The larger value of SR represents the more salient of the block. Considering temporal characteristics, the final salient map SRM, as demonstrated in Fig. 5., is then obtained by accumulating the SR in the consecutive video frames.



(a) **(b)**

Fig. 5. The red blocks in (a) denote a bag of visual salient cubes $\boldsymbol{W}$ and the SRM based on accumulating w is shown in (b)

When SRM is obtained, the carving surface can then be determined. For a vertical seam removal, the dynamic programming memorization table entry $MP(x, y)$ is given by

$$MP(x,y) = E_g(x,y) + \min \begin{cases} SRM(x-1,y-1) + MP(x-1,y-1) \\ SRM(x,y-1) + MP(x,y-1) \\ SRM(x+1,y-1) + MP(x+1,y-1) \end{cases}, \tag{5}$$

where $E_g(x, y)$ is the gradient energy of the position (x, y) in the current frame. The globally minimum energy seam is found by backtracking from the minimum value of the last row in MP to the first row. Using this ap-

proach, the chosen carving paths are determined according to the position of the bag of visual salient cubes and applied throughout a video sequence, as demonstrated in Fig. 6. Therefore, we can preserve most visual continuity of the salient cubes when the video sequence is resized by seam carving. Fig. 7(c) and 7(g) demonstrate that most carving paths can skip the salient cubes to protect the continuous visual dynamics using our proposed approach.

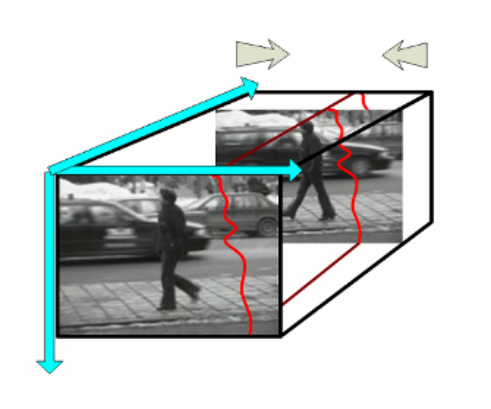


Fig. 6. The carving path determined by a bag of visual salient cubes is applied throughout the whole video sequence so that we can assure that most salient regions can be preserved

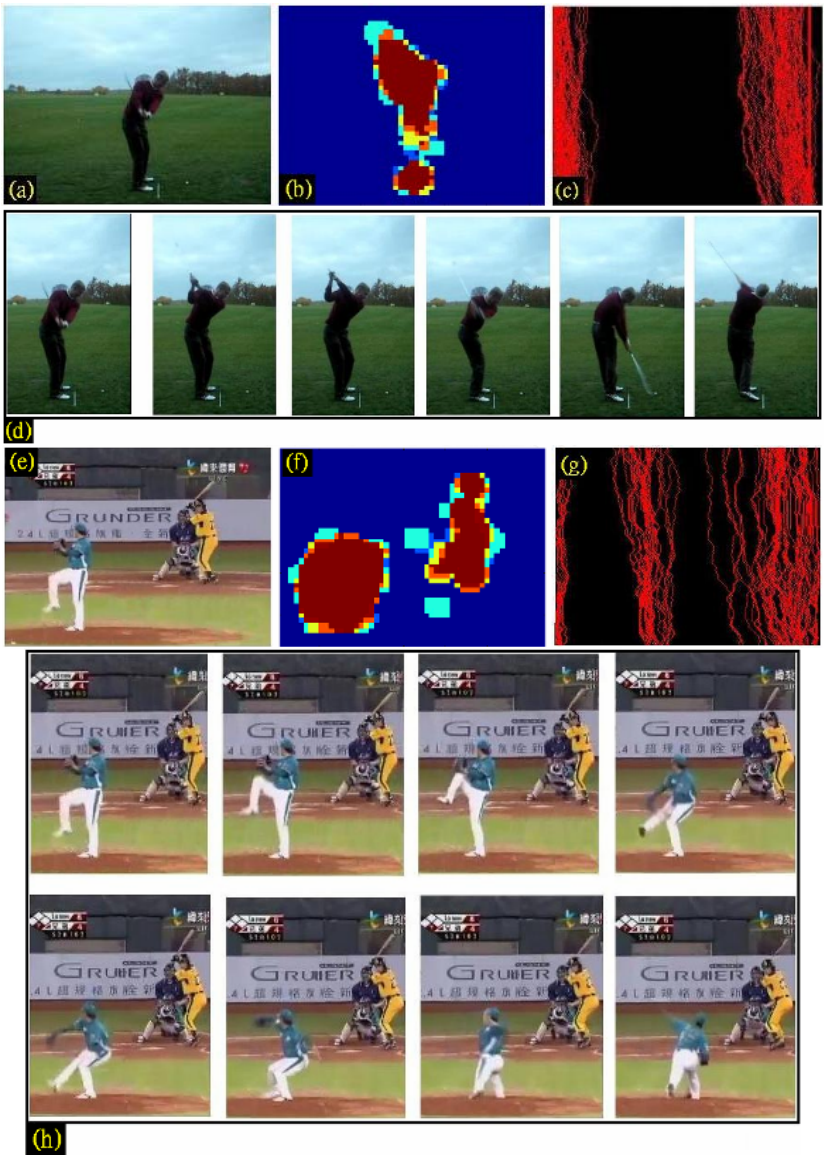


Fig. 7. (a)(e)original frames; (c)(g)carving paths determined based on (b)(f); (d)(h)the resulting consecutive carved frames

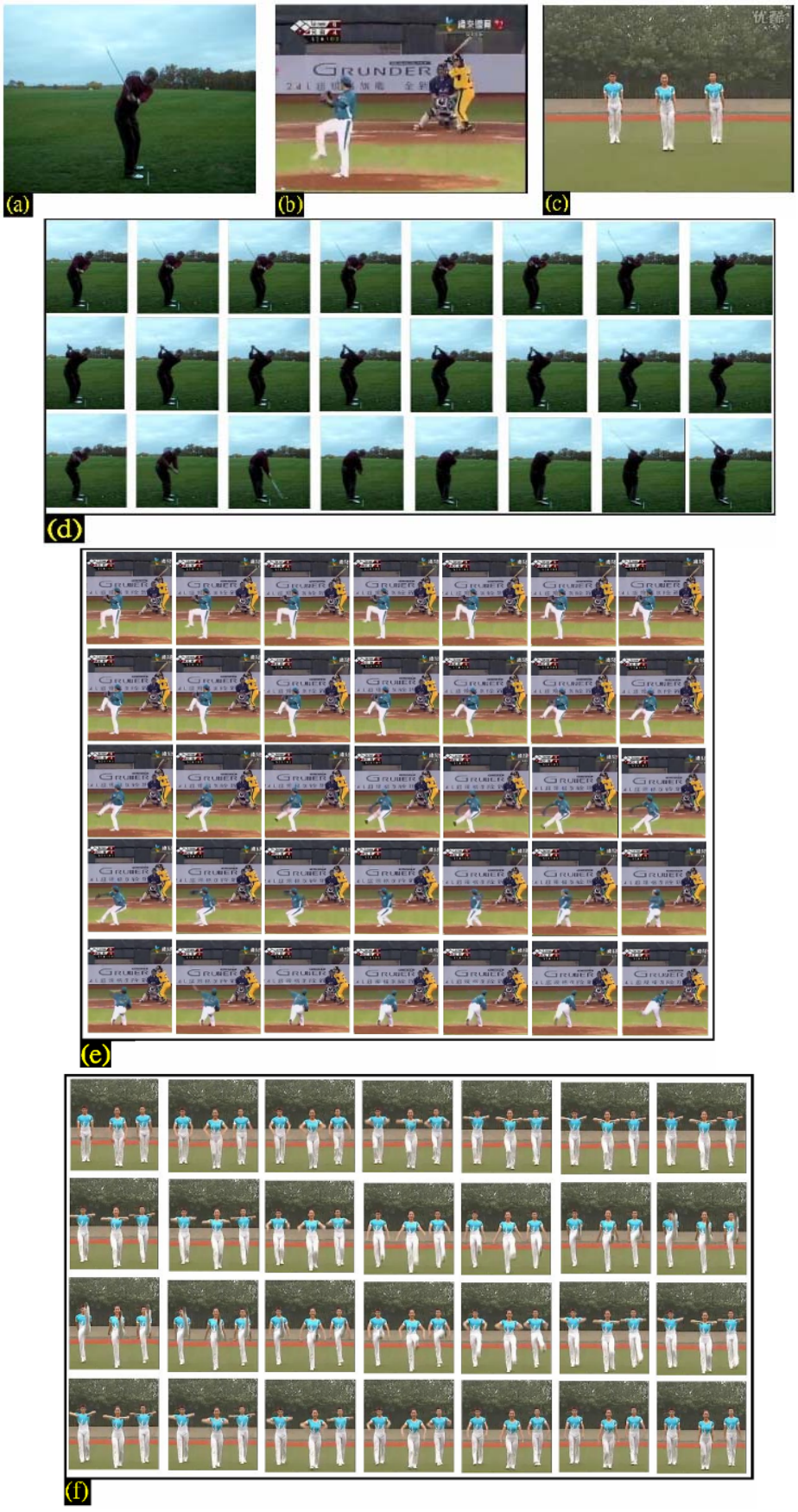


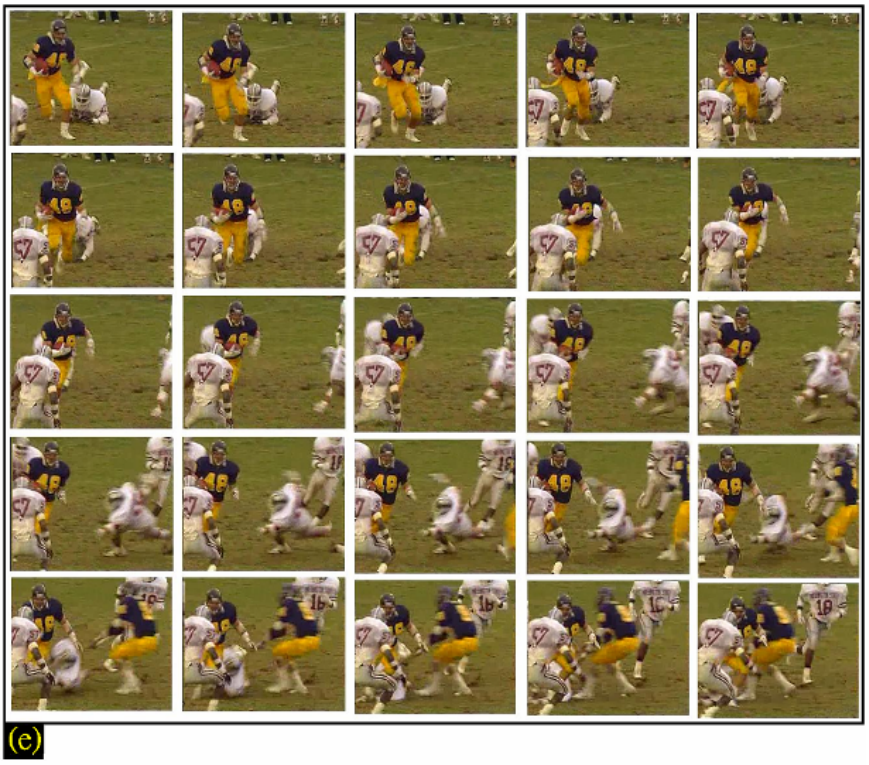
Fig. 8. (a)(b)(c) original video frames of resolution 320×240; (d) a golf playing sequence with a single moving target is resized to 170×240; (e) a baseball playing sequence with two salient targets is reduced to size 198×240; (f)a gym sequence with three salient targets is resize as 150×150

3 Experiment Results

To evaluate the performance of the proposed visual saliency model, we conducted experiments on different kinds of videos of size 320 × 240. The test videos included both standard MPEG-4 testing videos and a variety of videos downloaded from the Internet. The extensive dataset includes many kinds of videos with single salient object, multiple salient objects and dynamic backgrounds. All experiments are running with $\sigma = 1.25$ in Eq. (1), which is empirically determined. We collect 50 frames for each test sequence ($\nu = 50$) to build the spatiotemporal saliency map. The performance is measured by qualitative and quantitative evaluations in 3.1 and 3.2, respectively.

3.1 Qualitative Evaluations

Fig.8(a)-(c) show the testing videos that are of relatively static background. Their resulting carved videos are shown in Fig.8(d)-(f), respectively, in which the frames are in raster scan order. It can be observed that videos with single or multiple salient objects are well resized, in which the visual continuity of moving objects is successfully preserved. In Fig. 8(e), the pitcher and the hitter are both kept throughout the video sequence even the background text with high gradient energy have been removed. The effect of a bag of visual salient cubes results in protecting the salient dynamic regions. Fig.9(a)-(c) show the testing videos that are of dynamic background. Their resulting carved videos are shown in Fig.9(d)-(f), respectively. It can be observed that videos with single or multiple salient objects are well resized, in which the visual continuity of moving objects is successfully preserved. In Fig.9(d), the pedestrian is successfully detected and regarded as the most salient region and thus is completely preserved throughout the whole video sequence. In Fig.9(f), the tennis player in the resized video is mostly kept. Even small portion of his torso is removed; we can still watch and realize the content clearly.



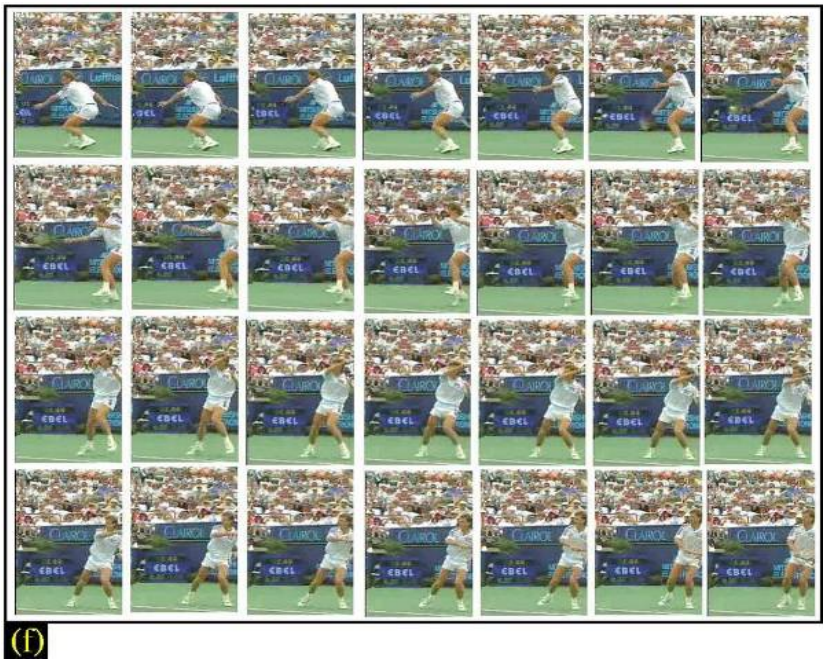
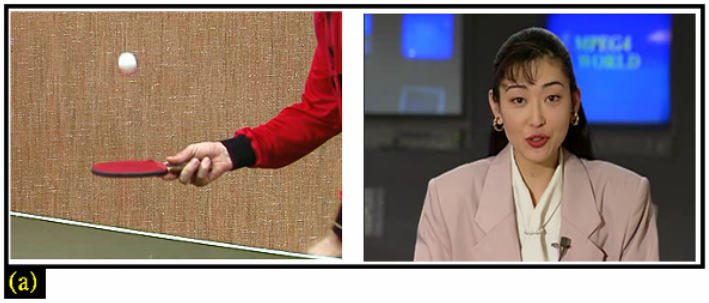
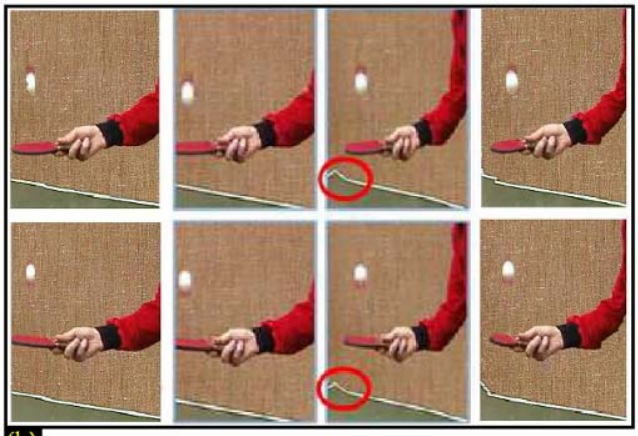


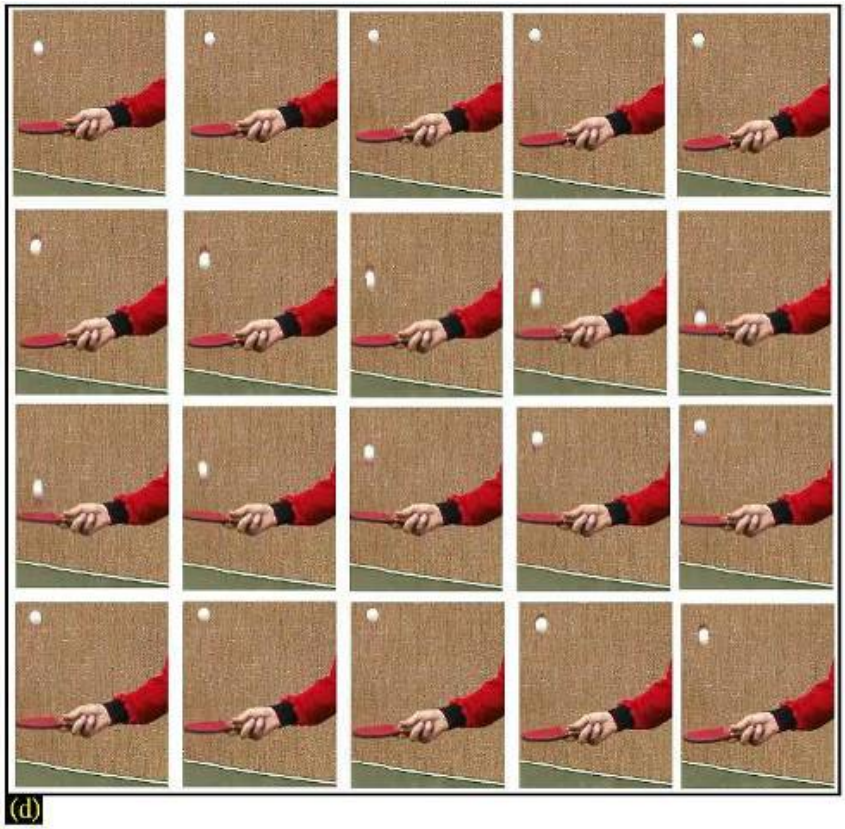
Fig. 9. (a)(b)(c) Original video frames with resolution 320×240; (a) a pedestrian walking sequence with relatively huge dynamic background. Its resized sequence of size 170×240 is shown in (d); (b) a MPEG-4 testing sequence "football" is of multiple moving targets with dynamic background. Its resulting resized video of size 250×200 is shown in (e); (c) a MPEG-4 testing sequence "Stefan". Its resulting video sequence of size 150×220 is illustrated in (f)

For further performance evaluation, we compare our proposed approach with related works of Hua et al. [8] and Wolf [17] using the MPEG-4 test dataset "Akiy" and "Tennis". The results are shown in Fig. 10. In Fig.10(b), we can observe that our approach outperforms the other three approaches since we can keep their isotropic manipulation and preserve most of the continuous dynamics of visual perception. The resized video sequences obtained by the methods in [8] and [17] both lead to the serious discontinuity in the table.


(a)


(b)


(c)


(d)

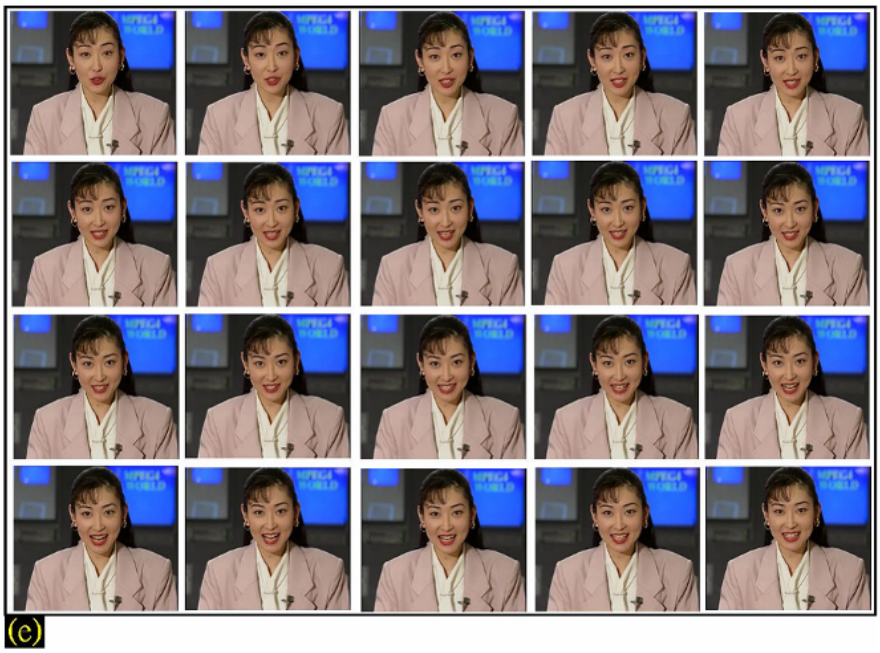


Fig. 10. (a) MPEG-4 testing videos of resolution 320×240, "Akiy" and "Tennis"; (b) From left to right: first column–our approach, second column–Hua's method [8], third column– Wolf et al.'s method [17] (by courtesy) and fourth column–direct scaling. (c) From left to right: first column–our approach, second column–direct scaling, third column–Hua's method [8] and fourth column–Wolf et al.'s method [17]. (d)(e) The resulting video sequences

3.2 Qualitative Evaluations

In subjective evaluations, we 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 resize results on 8 video clips from the proposed approach, and those from the direct scaling. Users would have five scores 1-5 to evaluate the performance where "1" is the worst and "5" denotes the best. There are 35 users with various backgrounds who participated in our user study. We first present the distribution of all the scores over the 8 clips from all the 35 users in Fig. 11. Over all the scores, 8.57% is poor, 14.3% is moderately, 25.7% is fair, 31.4% is good, and 20% is excellent.

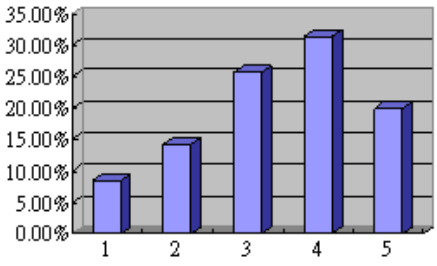


Fig. 11. The distribution of all the scores given by 35 users on 8 video clips. A score of 5 (1) is excellent (poor) about our approach

In objective evaluations, to quantify the performance of video seam carving, we use the PSNR (peak signal to noise ratio) for measurement. The PSNR is defined as

$$PSNR = 10 \times \log\left(\frac{255^2}{MSE}\right), \tag{6}$$

$$\text{where } MSE = \frac{\sum_{q=1}^{FrameSize} \left(I_q - P_q\right)^2}{FrameSize}. \tag{7}$$

MSE represents the mean square error, I_q is the value of the q_{th} pixel in the original frame and P_q denotes the q_{th} pixel in the processed frame. To compare the resulting video frames with the original ones, we manually label each salient region with a bounding box in the original frames and crop them through the sequence based on the initial bounding box. Fig.12. shows the average PSNR using the dataset in Fig.8(a) and Fig.10(a). In Fig. 12(a) although the PSNR altered and not stable from the beginning to the end, we can observe that in average the PSNR is over 49 which mean that the resulting quality of the resized videos is good enough to represent their regions of interest of the original video frames. The vibration of PSNR in Fig.12(b) is relatively large since the background is more complicated than that in the "Tennis" sequence.

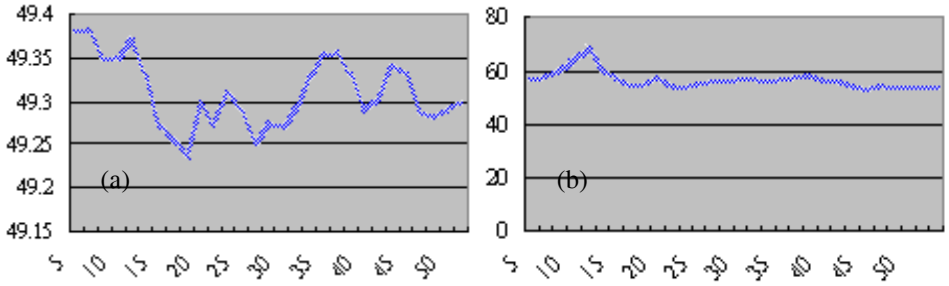


Fig. 12. (a) PSNR obtained using the dataset in Fig.8(a). (b) PSNR obtained using the dataset in Fig.10(a). The horizontal and vertical axes represent the frame number and the PSNR, respectively

4 Conclusion

Producing an appropriate extent of salient regions in video sequences by analyzing spatiotemporal visual attention is still a challenging problem. In this paper, we propose a novel approach for modelling dynamic visual attention based on spatiotemporal analysis in order to detect the focus of interest automatically. The continuously varied co-sited blocks in a video cube are first detected and their variations are characterized as a bag of visual cubes, which are further employed to determine a proper extent of salient regions in video frames. Once the proper extent through video cubes is determined, the carving process then can be conducted to find the global optimum. Our experiment shows that the proposed content-aware video seam carving based on spatiotemporal bag of visual cubes can effectively generate resized videos while keeping their isotropic manipulation and the continuous dynamics of visual perception.

References

[1] S. Avidan and A. Shamir, "Seam Carving for Content-Aware Image Resizing," *ACM Transactions on Graphics*, Vol. 26, No. 3, 2007.

[2] R. Achanta and S. Susstrunk, "Saliency Detection For Content-Aware Image Resizing," *Proceedings of the 16th IEEE International Conference on Image Processing*, pp. 1005-1008, 2009.

[3] W. James, *The Principles of Psychology*, Harvard Univ. Press, Cambridge, Massachusetts, 1980/1981.

[4] J. Yuan, Z. Liu, Y. Wu, "Discriminative Subvolume Search for Efficient Action Detection," *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.2442-2449, 2009.

[5] Y. Ke, R. Sukthankar, M. Hebert. "Event Detection in Crowded Videos," *Proceedings of IEEE International Conference on Computer Vision*, pp. 1-8, 2007.

[6] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues," *Proceedings of ACM International Conference on Multimedia*, pp.815-824, 2006.

[7] L. Laptev and T. Lindeberg, "Space-Time Interest Points," *Proceedings of IEEE International Conference on Computer Vision*, pp.432-439, 2003.

[8] G. Hua, C. Zhang, Z. Liu, Z. Zhang, Y. Shan, "Efficient Scale-Space Spatiotemporal Saliency Tracking for Distortion-Free Video Retargeting," *Proceedings of the 9th Asian Conference on Computer Vision*, pp. 182-192, 2009.

[9] F. Liu and M. Gleicher, "Video Retargeting: Automating Pan and Scan," *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 241-250, 2006.

[10] S. Li and M.C. Lee, "An Efficient Spatiotemporal Attention Model and its Application to Shot Matching," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.17, No.10, pp.1383-1387, 2007.

[11] Y.F. Ma, L. Lu, H.J. Zhang, M. Li, "A User Attention Model for Video Summarization," *Proceedings of ACM Multimedia*, pp.533-541, 2002.

[12] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, No.11, pp.1254-1259, 1998.

[13] V. Navalpakkam and L. Itti, "An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed," *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp. 2049-2056, 2006.

[14] L. Itti, and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, Vol.2, No. 3, pp. 194-203, 2001.

[15] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on In Systems, Man and Cybernetics*, Vol. 9, No. 1, pp. 62-66, 1979.

[16] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886-893, 2005.

[17] L. Wolf, M. Guttmann, D. Cohen-Or, "Non-homogeneous Content-driven Video-retargeting," *Proceedings of the 11th IEEE International Conference on Computer Vision*, pp. 1-6, 2007.