STEREOSCOPIC IMAGE RETARGETING BASED ON 3D SALIENCY DETECTION

Junle Wang[1,2], Yuming Fang[3], Manish Narwaria[1], Weisi Lin[4], Patrick Le Callet[1]

[1] LUNAM Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, France
[2]Ars Nova Systems, Nantes, France
[3]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China
[4] School of Computer Engineering, Nanyang Technological University, Singapore

ABSTRACT

In this paper, we propose a novel stereoscopic image retargeting algorithm based on 3D visual saliency detection. A new 3D visual attention model is designed based on 2D visual feature detection, depth feature detection and the modeling of various viewing bias in stereo vision. A geometrically consistent seam carving technique is adopted for retargeting stereo image pair. Experimental results demonstrated that both the proposed visual attention model and the proposed retargeting method outperform the state-of-the-art studies.

Index Terms— Visual attention, saliency detection, image retargeting, stereoscopic 3D, eye-tracking

1. INTRODUCTION

Stereoscopic 3D has recently received much attention as a result of a strong push from the cinema industry. Due to the emergence of 3D content (in cinema and at home) and 3D capable display devices, solutions to transfer the 3D materials from the theater to TV screen or various new devices supporting stereoscopic 3D display (e.g. tablets, smartphones) are expected.

During recent years, image/video retargeting algorithms have been widely used as an automatic solution of changing the size and aspect ratio of an image or a video with minimal visual distortion. These retargeting methods are content-aware. Generally, they can be differentiated by the following factors: (1) the methodology of computing the saliency map, which is used to measure the saliency of each image pixel; and (2) the methodology of image resizing. Compared with 2D image/video, stereoscopic 3D increases the sensation of presence through the enhancement of depth perception, and thus, the additional information (e.g. disparity) and constrains (e.g. issues related to binocular vision) raise serious challenges on both the saliency map detection step and the image resizing step.

In this paper, we propose a new image retargeting method which is particularly designed for stereoscopic images. The input to our method is a stereo image pair and a disparity map which can be either obtained by disparity estimation algorithms or depth range sensor. In order to predict the visual importance of different objects and regions in a scene, we first propose a novel visual attention model for stereoscopic images. The proposed model consists of the saliency detection based on 2D visual features and depth features. Additionally, we model various types of viewing bias during the viewing of 3D content. We integrate these viewing bias into the proposed model of 3D visual attention. Based on the predicted visual saliency map, we propose a retargeting algorithm to retarget the input stereo pair of image in the horizontal domain while minimizing both visual content distortion and depth perception distortion. Compared with other existing 3D retargeting methods, the geometric consistency of the scene can be better preserved by the proposed model.

2. BACKGROUND AND MOTIVATION

2.1 Stereoscopic image retargeting

Currently, various image/video retargeting algorithms having been proposed for image/video resizing [1]: content-aware cropping, scaling variants, rapid serial visual presentation, segmentation-based approach, seam-carving, warping-based method, path-based method, and multi-operator methods. However, most of these existing retargeting studies have focused on retargeting a single image or video. Regarding to retargeting stereo image pair, a relative small number of studies can be found in the literature.

An early work from Lang et al. [2] is a warping-based method focusing on adjusting the disparity map. The adjustment is conducted according to some stylistic considerations in order to adapt the depth range of the scene to the comfortable viewing zone of different displays and viewing conditions. However, their method does not actually resize the input image. Chang et al. [3] extended the method of [2] and proposed an approach which can both adapt the depth of scene and resize the stereoscopic images to the target aspect ratio. To detect salient regions in the scene, they apply a visual attention model which takes into account only 2D visual features but not the additional depth information provided by stereo vision. Lee et al. [4] proposed a retargeting approach which decomposes the input stereo images into multiple layers according to color and depth information. However, their approach also misses a validated 3D visual saliency model for detecting the salient region. Importantly, any discussion on geometric consistency of the scene is missing in all these three studies. Recently, Basha et al. [5] proposed an approach to generalize the single image seam carving algorithm to work on a rectified stereo pair. Their approach iteratively removes a pair of seams from the stereo image pair and takes into account the geometric consistency of the scene. However, a limitation of their approach is that they simply consider the close objects as salient objects as compared to further ones. No 3D visual attention model is applied in their approach.

2.2 Saliency detection for stereoscopic image

In order to determine which region of the input image can be modified while limit the visible artifact in the retargeted image,

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

most retargeting frameworks rely on visual attention models to detect the salient region in a scene. However, only a small number of stereoscopic visual attention models can be found in the literature. Most existing stereoscopic image retargeting algorithms, as introduced previously, still only rely on 2D visual attention models. However, studies have demonstrated that salient regions can result from depth features that are not considered in 2D visual attention models [6]. Therefore, it is not a trivial work to develop new visual attention model for 3D content and integrate the 3D visual attention model into the stereoscopic image retargeting framework.

During recent years, a few visual attention models [6] [7] [8] [9] have been proposed for 3D content. Some of them (e.g. [7] [8]) firstly use 2D visual features (e.g. color, intensity, orientation). The disparity map is then used as location prior knowledge in a later stage to weight the 2D saliency map. Most of these models rely on a simple assumption of viewing bias: the close region in a scene is always more salient than the further regions. However, recent studies have demonstrated the existence of other types of bias during the viewing of stereoscopic content, while few studies have taken into account these various bias in saliency detection. On the other hand, another type of 3D visual models considered the perception of depth as an additional and independent visual path (e.g. [6] [9]). These models focus on the extraction of depth features and the computation of a depth saliency map. The resulting depth saliency map is then pooled with 2D saliency maps to get the final 3D saliency map. However, these models do not take advantage of the various viewing bias [10] [11] in saliency detection. More importantly, the performance of most existing 3D visual attention models are not validated by eye-tracking data of stereoscopic images (except [6]). In this paper, we propose a hybrid model taking into account both 2D/depth features and various stereo viewing bias demonstrated in previous psychovisual studies. We validate the performance of the proposed model by a stereoscopic image eye-tracking dataset. The experimental results demonstrate better performance of the proposed model over relevant existing ones.

3. FRAMEWORK

In this section, we propose a stereoscopic image retargeting framework taking a pair of stereo image and the disparity map as input. The proposed framework consists of (1) a novel 3D visual attention model that takes into account 2D visual features, depth visual features and various viewing bias; and (2) an image retargeting method that deals with stereo image pair.

3.1. 2D saliency detection

Our previous study has demonstrate that DCT coefficients can be effectively used for feature extraction in saliency detection [12]. Similarly with our previous study [12], we extract the 2D saliency map for stereoscopic images based on DCT coefficients. Given an image, we first convert it into YCbCr color space and divide the image into small patches. Four features are extracted from each patch, including one luminance feature L (DC value of the Y component), two color features C_1 and C_2 (DC values of the Cb and Cr components), and one texture feature T (total AC energy of the Y component). The saliency value S_i^k for path i based on the feature k can be calculated as [9]:

$$S_i^k = \sum_{j \neq i} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d_{ij}^2}{2\sigma^2}} D_{ij}^k \quad (1)$$

where σ is the parameter for the Gaussian model; d_{ij} is the Euclidean distance between DCT block i and j; D_{ij}^k is the difference between blocks i and j with feature k.

Four feature maps can be calculated based on the Eq. (1) from features L, C1, C2, and T. The final 2D saliency map can be computed as:

$$S = \frac{1}{K} N(\mathrm{S}^k) \quad (2)$$

where N is the normalization operator and K is the number of feature maps (K=4).

3.2. Depth saliency detection

The first step of detecting depth saliency is to compute the perceived depth of each pixel based on its disparity, while this step is missing in most existing 3D visual saliency models.

The perceive depth D can be calculated by the following equation:

$$D = V/(1 + \frac{I \cdot R_x}{P \cdot W}) \quad (3)$$

where V represents the viewing distance between observer and screen plane, I represents the interocular distance, P is the disparity in pixels, W and R_x represent the width (in cm) and the horizontal resolution screen, respectively.

It has been demonstrated that depth contrast is a dominant feature in depth perception [13]. In this study, we adopt the approach in [6] to compute the depth saliency based on local depth contrast. A Difference of Gaussian (DoG) filter is first applied on the perceived depth map for extracting depth contrast. The depth saliency map is then computed based on the definition that the depth saliency S of each pixel equals the probability of this pixel being gazed at, given the depth contrast of this pixel: $S = P(C = 1 \mid f_{contrast})$. This probability function has been modeled in the study of Wang et al. [6]. Therefore, once a depth contrast map is computed, the depth saliency map can be obtained.

3.3. Modeling of viewing bias

Previous studies have demonstrated the 'center-bias' (or 'central fixation bias'): observers pay more attention to the regions closer to the center of the screen. Due to a large impact of image borders on center-bias [10] [14], we apply an anisotropic 2D Gaussian function to model the 2D center-bias:

$$W_{2D\_C}(x, y) = \exp\left(-\frac{(x - x_0)^2}{2\sigma_x^2} - \frac{(y - y_0)^2}{2\sigma_y^2}\right) \quad (4)$$

where (x_0, y_0) represents the coordinates of the screen center, σ_x and σ_y denote the standard deviations in the horizontal and vertical direction, respectively. The two standard deviations are computed as:

$$\sigma_y = \sigma_x \times \left(\frac{R_x}{R_y} Ind(R_x < R_y) + \frac{R_y}{R_x} Ind(R_x > R_y)\right) \quad (5)$$

where R_x and R_y are the width and height of the image, and *Ind()* is the indicatric function. In order to take into account the viewing distance in modeling 2D center-bias, σ_x and σ_y are measured in visual degree.

One reason for 2D center-bias is that people can save effort of moving their gaze point to most region of the scene if they start their observation from the center of the scene and move their gaze back to the center after the observation. Moreover, compared to other locations, looking at the center also allows observers to acquire more information about the scene [10]. Therefore, it is

reasonable to assume that people's attention is not only bias to the screen center but also bias to the middle depth plane of the scene in stereo viewing condition. We here define the middle depth d_m plane as:

$$d_m = (d_{\max} + d_{\min})/2 \tag{6}$$

where d_{max} and d_{min} represent the maximum and minimum disparity of the scene. This bias to the middle depth plane may save the effort of vergence change when people converge their eyes to various depth planes in the scene. Additionally, due to the limitations of human's vision (e.g. the Panum's area and depth-of-field), looking at the middle depth plane allows an accurate vision on more objects that locate at different depth planes, since the diplopia (i.e. double vision) and defocus blur can be limited. In this study, we named this bias as 'central depth bias', and model it by a Gaussian function taking into account the depth difference between each pixel and the the middle depth plane:

$$W_{Central\_D}(x,y) = \exp\left(-\frac{(d(x,y)-d_m)^2}{2\sigma_{d_1}^2}\right) \tag{7}$$

where $d(x,y)$ represents the disparity value of each pixel. The standard deviation, σ_{d1} , is a function of the depth range of the whole scene.

Moreover, another viewing bias to the close object in stereo viewing condition has been widely used in previous studies. We thus called it as 'front depth bias'. A recent study [11] demonstrates that the front depth bias concerns only the relative depth but not the absolute depth (i.e. the absolute disparity). So the saliency of each pixel is related to its depth difference from the region having the smallest distance to the observer. We model the front depth bias by a Gaussian function:

$$W_{Front\_D}(x,y) = \exp\left(-\frac{(d(x,y)-d_{\min})^2}{2\sigma_{d_2}^2}\right) \tag{8}$$

where d_{min} denotes the minimum depth value (i.e. the pixel closest to the observer) of the scene. The standard deviation, σ_{d2} , is a function of the depth range of the whole scene.

3.4. Saliency pooling and weighting

We adopt a straightforward approach to merge the depth saliency map (SM_{dep}) with the 2D saliency map (SM_{2D}) which is based on a linear combination. After the depth saliency map and the 2D saliency map are combined, we then weight the resulting saliency map by the three viewing bias.

$$SM_{final} = (\omega_1 SM_{2D} + \omega_2 SM_{dep}) \cdot \prod W_\theta \tag{9}$$

where, $\omega_1 = 0.5$, $\omega_2 = 0.5$ and $\theta \in$ (2D_C, Central_D, Front_D).

3.5. Gradient based energy map

In addition to visual saliency map, energy maps are also considered to limit the visual distortion in the retargeted image. As proposed in [5], the energy function of the stereo seam carving consists of terms related to intensity, depth and correspondence between the seam pair in the left and right images:

Figure 1: Examples of the salient region detection. First row to the final row: original images; saliency map from the proposed model; saliency map from eye-tracking data.

$$E_{stereo} = E_{intensity} + E_{depth} + E_{correspondence} \tag{10}$$

3.5.1. Intensity energy

The intensity energy is computed for both horizontal and vertical directions, each of which is computed as followed:

$$E_{intensity} = E_{horizontal} + E_{vertical} \tag{11}$$

$$E_{horizontal} = |I(x-1,y) - I(x+1,y)| \tag{12}$$

$$E_{vertical} = \begin{cases} \sum_{k=x'+1}^{x} |I(k,y-1) - I(k-1,y)| & x' < x \\ 0 & x' = x \\ \sum_{k=x+1}^{x'} |I(k-1,y-1) - I(k,y)| & x' > x \end{cases} \tag{13}$$

where x' denotes the seam's pixel removed from the original image.

3.5.2. Depth energy

The depth energy term, E_{depth}, is taken into account for the purpose of minimizing the depth distortion. The computation of the depth energy is similar to the horizontal intensity energy term, except that the intensity function I in Eq. (13) is replaced by the disparity map.

3.5.3. Correspondence energy

Nowadays, generating a precise disparity map is still a challenging topic. In order to reduce the influence of errors in the estimated disparity map, it would be better to remove pixels with high confidence of disparity values during the seam carving. As in [5], we define the correspondence energy as the difference in the intensities of corresponding pixels in the image pair:

$$E_{correspondence} = |I_{Left}(x,y) - I_{Right}(x - D(x,y), y)| \tag{14}$$

3.6. Seam selection and carving

In this study, seam carving is performed for stereoscopic images, so the visibility relations between corresponding pixels in the stereo image pairs need to be taken into account. Previous studies [5] have shown that the occluded pixels (i.e. pixels that are visible in only one view) and the occluding pixels (i.e. the pixels that occlude the occluded pixels) should not be removed. Therefore, in our method, discontinuous seams are allowed though continuous seams are preferred by default.

We use the approach proposed in [5] to perform the piecewise seam selection and carving based on the stereoscopic visual

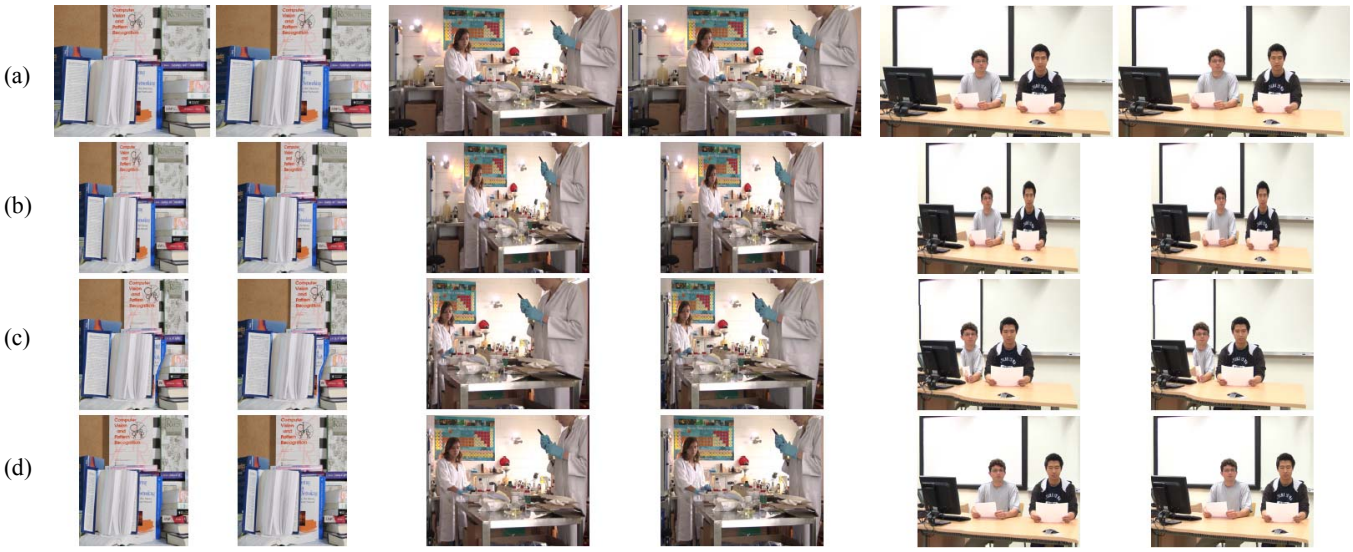

Figure 2: Comparison of different image retargeting methods. (a) The left/right view of the original stereo image pair. (b) - (d) Retargeted images from linear rescaling method, Basha's method [5], and the proposed method.

saliency map, energy map, and the visibility of pixels. Formally, the optimal seams are selected using dynamic programming:

$$M(x,y)=\begin{cases}\min\limits_{x'\in\{x-1,x,x+1\}}(S(x,y)+E_{stereo}(x,y,x')) & B(x,y)=0\\ \min\limits_{x'\in[m]}(S(x,y)+E_{stereo}(x,y,x')) & B(x,y)=1\end{cases} \quad (15)$$

where M is a cost matrix determining the pixel position (x,y) for the optimal seams; $B(x,y)$ indicates whether a continuous path is blocked in row y-1 by a occluded/occluding pixel. Note that each selected seam in the left image is matched to a seam in the right image. The coupled seams in both images are removed simultaneously.

4. RESULTS

We firstly tested our visual saliency model and the seam carving method on the IVC 3DGaze dataset [6]. To the best of our knowledge, this is the only available stereoscopic image database which provides eye tracking data. This database includes various types of images (e.g. indoor scene, outdoor scene, and scenes containing various numbers of objects). We use all the 18 stereoscopic image for the evaluation. Ten of the images come from the Middlebury Stereo Datasets [15] which has been widely used for performance evaluation in previous studies of stereoscopic image retargeting.

Some examples of the saliency maps predicted by the proposed visual attention model are presented in Figure 2 for a qualitative evaluation. To quantitatively evaluate the model's performance, we apply the same metrics as in [6]: (1) the Pearson Linear Correlation Coefficient (PLCC), (2) the Kullback-Leibler divergence (KLD), and (3) the Area Under Receiver Operating characteristics Curve (AUC). The first two metrics are directly applicable to a comparison between a ground-truth fixation density map and a predicted saliency map, whereas AUC is usually applied to compare the actual fixation points to a predicted saliency map. To the best of our knowledge, [6] is currently the only study that evaluates the performance of 3D visual attention models by eye-tracking data of stereoscopic images. We thus compare the performance of the proposed model with the three models introduced in [6]. The result (see Table 1) shows that the proposed model can predict more accurately the saliency.

Table 1. Comparison results of PLCC, KLD, and AUC values from different stereoscopic saliency detection models.

Models	PLCC	KLD	AUC
Model 1 in [6]	0.356	0.704	0.656
Model 2 in [6]	0.424	0.617	0.675
Model 3 in [6]	0.410	0.605	0.670
The proposed model	**0.652**	**0.300**	**0.720**

In order to evaluate the performance of our retargeting method, we compare our proposed method with (1) a simple rescaling method and (2) a state-of-the-art stereo image retargeting algorithms recently proposed by Basha et al. [5]. For the linear rescaling method, we applied the same linear scaling operator on both the left and right images to keep the geometric consistency. Moreover, to the best of our knowledge, Basha's method is the only stereo image retargeting method that takes into account the geometric consistency of the scene's depth information.

In the experiment, the input stereo image pairs are resized to 70% of its original width. The height of the input image is unaltered. Since there is not yet reliable objective/subjective methodologies for quantitatively evaluating the visual quality of retargeted image, we do a qualitative evaluation of the retargeting results in our study. In Figure 2, we can see that objects in the retargeted images created by linear rescaling and Basha's method are obviously distorted as compared to the original images. Particularly, from the second and third example, we can see that the people locating at the far part of the scene are largely distorted. This distortion results from the lack of an efficient visual attention model in Basha's method, which just considers the near objects to be salient. We can also see that the salient objects are preserved well in our proposed method (i.e. the last row in Figure 2).

5. CONCLUSION

In this paper, we have proposed a novel stereoscopic image retargeting method based on 3D visual saliency detection and stereo seam carving, for the purpose of detecting 3D salient regions and preserving geometric consistency of the 3D scene. We have demonstrated, qualitatively and quantitatively, that the effectiveness of our novel saliency detection model as well as good performance of the proposed seam carving algorithm for stereoscopic image pair retargeting.

6. REFERENCES

[1] D. Vaquero, M. Turk, K. Pulli, M. Tico, and N. Gelfand, "A survey of image retargeting techniques," 2010, vol. 7798, pp. 779814–779814–15.

[2] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Trans Graph*, vol. 29, no. 4, pp. 75:1–75:10, Jul. 2010.

[3] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang, "Content-Aware Display Adaptation and Interactive Editing for Stereoscopic Images," *IEEE Trans. Multimed.*, vol. 13, no. 4, pp. 589–601, 2011.

[4] K.-Y. Lee, C.-D. Chung, and Y.-Y. Chuang, "Scene warping: Layer-based stereoscopic image resizing," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 49–56.

[5] T. Basha, Y. Moses, and S. Avidan, "Stereo Seam Carving A Geometrically Consistent Approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. Early Access Online, 2013.

[6] J. Wang, M. P. DaSilva, P. LeCallet, and V. Ricordel, "A Computational Model of Stereoscopic 3D Visual Saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151–2165, 2013.

[7] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic Visual Attention Model for 3D Video," in *Advances in Multimedia Modeling*, S. Boll, Q. Tian, L. Zhang, Z. Zhang, and Y.-P. P. Chen, Eds. Springer Berlin Heidelberg, 2010, pp. 314–324.

[8] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, "Adaptive 3d rendering based on region-of-interest," in *IS&T/SPIE Electronic Imaging*, 2010, p. 75240V–75240V.

[9] E. Potapova, M. Zillich, and M. Vincze, "Learning What Matters: Combining Probabilistic Models of 2D and 3D Saliency Cues," in *Computer Vision Systems*, J. L. Crowley, B. A. Draper, and M. Thonnat, Eds. Springer Berlin Heidelberg, 2011, pp. 132–142.

[10] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, Jul. 2009.

[11] J. Wang, P. Le Callet, S. Tourancheau, V. Ricordel, and M. Perreira Da Silva, "Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli," *J. Eye Mov. Res.*, vol. 5, no. 5, pp. pp. 1–11, Sep. 2012.

[12] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency Detection in the Compressed Domain for Adaptive Image Retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, 2012.

[13] A. Brookes and K. A. Stevens, "The analogy between stereo depth and brightness," *Perception*, vol. 18, no. 5, pp. 601–614, 1989.

[14] J. Wang, M. Perreira Da Silva, P. Le Callet, and V. Ricordel, "Study of center-bias in the viewing of stereoscopic image and a framework for extending 2D visual attention models to 3D," 2013, vol. 8651, pp. 865114–865114–9.

[15] D. Scharstein and C. Pal, "Learning Conditional Random Fields for Stereo," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, 2007, pp. 1–8.