AUTOMATIC STEREOSCOPIC 3D VIDEO REFRAMING

Lino Coria, Di Xu, Panos Nasiopoulos

The University of British Columbia
Vancouver, BC, Canada
{linoc, dixu, panos}@ece.ubc.ca

ABSTRACT

3D displays have various aspect ratios (e.g., 16:9, 4:3, and 3:2). Watching 3D videos with the wrong aspect ratio decreases the quality of the viewing experience. We have developed a smart reframing solution that uses a visual attention model for stereoscopic 3D video to identify the prominent visual regions of every stereoscopic frame. Our method uses several saliency indicators such as depth, edges, brightness, color, and movement. Additionally, our method provides a dynamic cropping window that slides smoothly from frame to frame.

Index Terms — 3D, stereoscopic, 3D TV, 3D displays, aspect ratio, reframing, visual attention model, saliency map, quaternion, depth map, texture map.

1. INTRODUCTION

3D-capable consumer devices are currently available in many different aspect ratios. 3D TVs, usually larger than 40", feature a 16:9 aspect ratio while it is common for smaller displays such as the ones found on tablets and mobile devices to have 4:3 and 3:2 aspect ratios, respectively (see Fig. 1). Stereoscopic 3D media creators tailor their content for a specific aspect ratio (usually 16:9). Unfortunately, playing this content on 3D displays with aspect ratios that are different to the intended one might degrade the quality of the viewing experience.

Fig. 1. Three different aspect ratios.

Several solutions have been proposed in order to compensate for this variation in aspect ratios. The straightforward option is to add black bars to the screen, which can be horizontal (also known as *letterboxing*) or vertical (also known as *pillarboxing*), depending on the original and new aspect ratios [1]. The main problem with this option is that a significant part of the screen will remain unused. This is particularly problematic for small devices. A second option consists of cropping the borders of the video frames so that the modified frames have the proper aspect ratio. This technique, known as centered cropping, eliminates visual information without taking into account that these regions might actually be of interest to the viewers.

An alternative solution is to supervise the reframing process and manually choose the areas of interest on a frame-by-frame basis. Human observers can detect the important visual points on the screen and control the location of the bounding box (i.e., the region of the frame that will prevail after the reframing process). This process, known as pan and scan in 2D video, ensures that the modified content will be meaningful to the viewers. This is evidently expensive, time-consuming and not suitable for real-time applications. A better solution is to have an automatic process that identifies the main visual information and keeps it inside the bounding box.

A number of methods have been proposed for automatic content reframing. A vast majority of these schemes, however, deal exclusively with 2D still images [2], [3], [4]. For the case of automatic 2D video reframing, [5] proposes a scheme that preserves visually important regions as well as temporal stability. The Visual Attention Model (VAM) used in this scheme is taken from [6], [7]. Two Kalman filters are used to ensure good temporal consistency by smoothing the change in the values of the bounding box center coordinates on every frame.

Color and depth are employed to create a visual attention model for 3D images in [8]. Results and conclusions, however, were drawn using data from merely five stereoscopic images.

An early proposal for a visual attention model for 3D video is found in [9]. The proposed scheme uses cues such as stereo disparity, image flow and motion. Relative depth was employed as a target selection criterion. This scheme is able to detect the moving object that is closest to the cameras. Although this solution might be useful for some videos, it will not provide acceptable results for complex scenes like the ones usually found in commercial videos made by the entertainment industry.

A VAM for 3D video is presented in [10]. The model uses features such as depth information, luminance, color and motion. This scheme, however, was developed and tested on multi-view videos and cannot be directly applied to stereoscopic video content. It is also computationally expensive, making it unsuitable for near real-time application.

Most of the schemes proposed in the literature provide a saliency map as the end result. For automatic video reframing, however, this is only an intermediate step towards a final solution. Once the saliency data has been obtained, a decision has to be made as to what sections of each frame need to be cropped. This is particularly challenging for the case of video since the bounding box cannot change abruptly from frame to frame. The case of 3D video adds an extra challenge: careless cropping might cause window violations that will produce an unpleasant 3D experience to the viewers. Having these issues in mind, we have developed a complete solution that identifies the prominent visual regions by using a VAM for stereoscopic 3D video. Our method uses several saliency indicators such as depth, edges, brightness, color, and movement. Additionally, our method provides a dynamic bounding box that slides smoothly from frame to frame and lowers the chances of getting window violations.

The rest of the paper is organized as follows. Section II provides an introduction to our visual attention model and the creation of depth maps. Section III describes our algorithm, including the choice of the bounding box and the temporal smoothing process. Section IV shows our experiments. Conclusions are drawn in Section V.

2. OUR 3D VISUAL ATTENTION MODEL

We have developed a Visual Attention Model (VAM) for 3D content that computes a disparity saliency map and combines this information with a couple of 2D saliency indicators, namely, local edges and global texture. Local edges emphasize the boundary of the objects in an image or video frame while global texture saliency refers to basic visual features that attract people's attention such as color, brightness motion. Disparity-based saliency assumes that objects that are close to the camera draw more visual attention than distant objects. This information can be obtained by comparing the left and right views of a stereoscopic image. Fig. 2 provides an example of how our scheme combines the 3D disparity data with the two other maps to produce a definitive saliency map for 3D video.

2.1 Local Edge Saliency Map

We compute a local edge saliency map for each frame. This map is computed as the gradient of the intensity values of the frame's pixels. An example is shown in Figure 2b.

2.2 Disparity Saliency Map

People tend to give more importance to objects that are closer to them than to the ones that are further back. Information about the closeness of objects can be obtained by comparing the left and right views of each frame. The steps for computing a disparity-based saliency map are as follows.

In order to retain a fast algorithm, we first down-sample the left and right views of each stereoscopic frame. We then extract distinctive feature points from each down-sampled view employing the shift-invariant feature transform [11]. Feature points from both views are matched, and disparities of each pair of matching points are computed. Next, we remove some pairs of the matching points to further ensure the matching accuracy of the remaining points. This is done by discarding the points with large vertical disparities and points with horizontal disparities that heavily deviate from the majority of the points. A pruning algorithm [12] is then used to retain a set of sparse feature points in order to further increase the robustness and accuracy. This pruning algorithm trims the less robust points based on their temporal stability. We match and track all feature points among temporal frames. A point is removed if it is temporally less robust and its disparity is similar to a more robust neighboring point.

Subsequently, a dense disparity map (one value per pixel) is generated by linearly interpolating the sparse feature points based on Delaunay triangulation. Areas that are not included in any Delaunay triangle are assigned the maximum disparity value. The disparity-based saliency map is finally obtained by assigning high saliency to small disparity values and low saliency to large disparity values. Refer to Fig. 2d for an example of this type of map.

2.3 Global Texture Saliency Map

It has been reported [13] that viewers pay special attention to basic visual features such as color, brightness and motion. Therefore, it is important to use this information to determine the salient objects of video frames. Although several computational models have been proposed to simulate human visual attention [14], we decided to use the scheme proposed in [15] as a starting point for our global texture map since it is fast and produces better results than other state-of-the-art schemes.

For every pixel, we express information related to color, intensity and motion in the form of a quaternion [16]. This allows us to obtain a quaternion frame q.

$$\mathbf{q} = \boldsymbol{S_d}\mathbf{M} + \alpha\mathbf{C}_1\mu_1 + \beta\mathbf{C_2}\mu_2 + \gamma\mathbf{I}\mu_3 \quad (1)$$

where $\boldsymbol{S_d}$ is the normalized disparity map; α, β, and γ are constant values between 0 and 1; μ_i, i = 1, 2, 3 satisfies $\mu_i^2 = -1, \mu_1 \perp \mu_2, \mu_2 \perp \mu_3, \mu_3 \perp \mu_1, \mu_3 = \mu_1\mu_2$. **M** is the motion channel, $\mathbf{C}_1$ and $\mathbf{C}_2$ are the two color channels recommended in [17] (red/green and blue/yellow, respectively), and **I** is the intensity channel. $\mathbf{C}_1$, $\mathbf{C}_2$ and **I** are computed as in [15]. In order to obtain **M**, we first compute the absolute value of the difference between the intensity channel of the current frame $\mathbf{I}(n)$ and the intensity channel from a previous frame $\mathbf{I}(n - n_0)$, where n_0 is a small positive integer number. The obtained result is normalized so that the highest value equals 1. For every pixel, the motion component takes the value of this normalized quantity provided that it is above a certain threshold τ_M, where $0 < \tau_M \leq 1$. Otherwise, the motion component for that pixel equals 0.

Readers familiar with [15] will notice that we have added weights to the sum of the four channels in (1). This decision was taken after implementing the original scheme and conducting several subjective tests with a small group of people and determining that, for 3D videos, the motion channel is more relevant than the other channels to create an effective visual attention model. As in [15], we also use the Quaternion Fourier Transform (QFT) [16] to produce a saliency map. We implemented this method using the quaternion toolbox for Matlab offered in [18].

An example of a global texture saliency map is shown in Fig. 2c. The map indicates that the most salient regions of the frame are the bright light seen through the window and the only section of the frame with significant movement, which is the right arm of the lady (the sequence is handheld so there is relative movement in the entire frame).

2.4 Combined 3D Saliency Map

Finally, we fuse the normalized local edge saliency map $\boldsymbol{S_l}$, the normalized disparity saliency map $\boldsymbol{S_d}$, and the normalized global texture saliency map $\boldsymbol{S_g}$ as a combined 3D saliency map $\boldsymbol{S_c}$, by computing the average value: $\boldsymbol{S_c} = (\boldsymbol{S_l} + \boldsymbol{S_d} + \boldsymbol{S_g})/\mathbf{3}$. (2)

3. AUTOMATIC 3D VIDEO REFRAMING

Our automatic stereoscopic 3D video reframing solution produces the three saliency maps described in Section II and fuses them to create a single model for visual attention. There are several proposals for combining saliency maps such as the schemes detailed in [3]. For our method, the maps are normalized and averaged to obtain the combined saliency map.

For the case of video, decisions on how to crop a frame so that it fits its new aspect ratio cannot be solely based on the information available from its associated saliency map. We also need to consider the cropping locations of the previous frames so that we can ensure that the location of the bounding box does not result in video flickering. To achieve this, we have designed a scheme that provides smooth temporal cropping.

The first goal of our scheme is to identify the area in the saliency map with the highest "energy." The energy in an area is defined as the summation of all saliency values within this area. For a fast implementation, an accumulated energy matrix is precomputed. We normalize the accumulated energy matrix so that the maximum value in the matrix is 1. The value of this matrix at each location P, denoted as E(P), is calculated as the energy of the rectangular region defined between the pixel on the top-left corner of the map and the current pixel P. Then, the energy in any rectangular region in the map can be later computed as three

summations rather than requiring the sum of all the pixel values in this area. As shown in Fig. 3, the energy in the rectangle ABCD can be simply computed as:

$$E(\text{ABCD}) = E(\text{A}) - E(\text{B}) - E(\text{D}) + E(\text{C}). \quad (3)$$

Based on the desired aspect ratio, we crop the frame leaving the rectangular region that contains the highest energy.

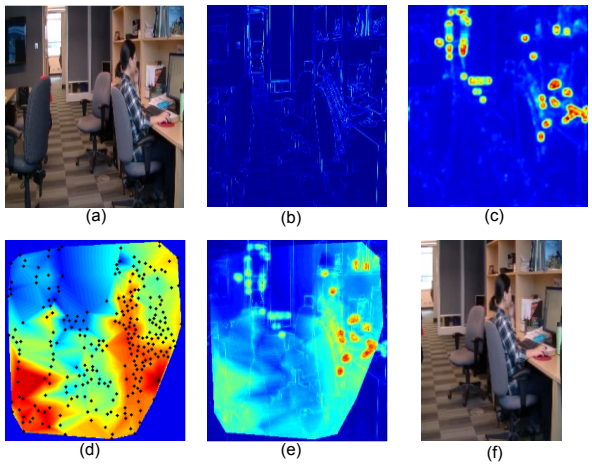

Figure 1. The proposed Visual Attention Model combines three saliency maps (only left frame is shown). (a) Original frame (vertically squeezed since it is the left half of the side-by-side 16:9 3D frame; (b) local edge saliency map; (c) global texture saliency map; (d) disparity saliency map; (e) combined map using the same weight for all three maps; (f) resulting bounding box with a 4:3 aspect ratio.

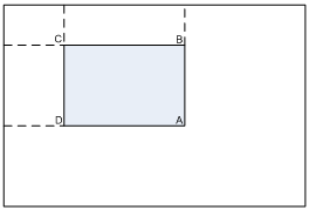

Figure 3. The energy in the rectangle ABCD is defined as E(ABCD) = E(A) - E(B) - E(D) + E(C).

Quite often, parts of an object have high saliency values whereas other parts have low values. Reframing solely based on the energy of a saliency map may result in cropping some important object. In order to avoid this, we propose to use a very simple yet effective approach. First, we slightly reduce the size of the bounding box by μ pixels when searching for the highest energy area. Then, we expand the bounding box by μ pixels on all sides with the purpose of including the entire important object in the cropped new frame. This shrinking and expanding approach also implicitly brings the salient area towards the center of the new frame. Furthermore, this scheme reduces the probability of experiencing window violation after reframing.

In order to ensure the temporal stability of the locations of the cropping window, we first make sure that the locations of the consecutive frames are spatially constrained if no scene change is detected. That is, the location difference of two consecutive frames is smaller than a threshold δ. The value of δ is determined by the resolution of the original video and the amount of motion contained in the sequence.

Although a constraint of cropping locations is set in the previous step, local jerks still exist. This is often caused by small differences on the consecutive saliency maps which result from insignificant motion or lighting changes. Therefore, when choosing the bounding box for the current frame, we give higher priority to the bounding box location of the previous frame. To this end, if the energy increase associated with the new location is less than a threshold τ_E, we keep using the previous location. In addition, a temporal filter using a simple moving average algorithm is employed to further ensure the smoothness of the cropping locations. Finally, the cropping locations are rounded to integers after applying the temporal filter in order to avoid spatial interpolation of a frame.

4. EXPERIMENTAL RESULTS

We captured dozens of HD (high definition) stereoscopic video sequences using a 3D video camera. Each video frame is composed of a side-by-side left and a right view, each with an 8:9 aspect ratio, resulting in a 3D frame with a 16:9 aspect ratio. This format is widely accepted by 3D displays of 16:9 aspect ratios. The video resolution of the side-by-side frame is 1920 pixels × 1080 pixels. The videos are several seconds long (from 10 to 43 seconds) and some of them were recorded with a handheld camera and others had the camera mounted on a tripod. Most of the videos feature people working, playing or walking and we include both indoor and outdoor sequences.

We reframed these videos to a 4:3 aspect ratio (i.e. a 2:3 aspect ratio for each view) using the proposed method. We used a value of 5 for n_0. The threshold τ_M was set to 0.6 and the weights α, β, and γ were all set to 0.1.

For HD stereoscopic video sequences, we employ μ =100 pixels in the shrinking and expanding stage. We found in our experiments that a δ value of 15 pixels is able to sufficiently track the moving objects and maintain a relatively constrained position of the bounding box. The energy increase threshold τ_E is set to 1%. Finally, a window size of 20 frames is used for the window of the smoothness filter.

Due to space limitations we will only provide two examples of how our method takes decisions based on both the combined saliency map and the positions of the neighboring bounding box.

Fig. 4 shows the various saliency maps created for one of the frames of a sequence called “Playground” which was captured with a handheld camera. In this example, both the global texture map and the disparity map indicate that the most relevant region of the frame is the young girl. The global texture map highlights her because she is moving while she is emphasized by the disparity map because she is close to the 3D camera. The combined saliency map clearly indicates that the young girl occupies the most salient region of the frame. This map informs our reframing method where to place the bounding box.

Another example is illustrated in Fig. 5 which includes a frame from the sequence “Main Mall,” captured with a 3D camera placed on a tripod. In this case, the global texture map highlights the people walking down the street. On the other hand, both the local edge map and the disparity map highlight the bicycle as the main object on the frame. Finally, the combined map identifies all the salient regions of the frame. This allows our method to select a bounding box for reframing purposes. Fig. 5 is a clear example of the importance of employing a vast number of features to identify the objects of highest visual interest for each stereoscopic video frame. The combination of the various saliency maps provides an accurate visual attention model for 3D content.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel and complete pipeline to reframe 3D content for displaying on screens of different aspect ratios. We first compute a bottom-up saliency map that was carefully fused from luminance, chrominance, motion, and disparity

information. Then, we develop an automatic reframing approach to crop content based on this saliency map. Special attention was paid to avoid the important objects being cropped or locating at the border of the new frame. Temporal jerkiness of the cropping window was also eliminated by our proposed method. The results showed that our proposed scheme is very effective, robust, simple, and computational efficient for a great variety of stereoscopic video sequences. It works well for 3D videos that are captured with a tripod or handheld, still or panning, indoor or outdoor, with slow or fast motion, simple or complex scene. Results from subjective tests that show that viewers prefer our reframing solution to the traditional centered cropping approach will be presented in a future paper. We are currently working on a real-time implementation of this algorithm.

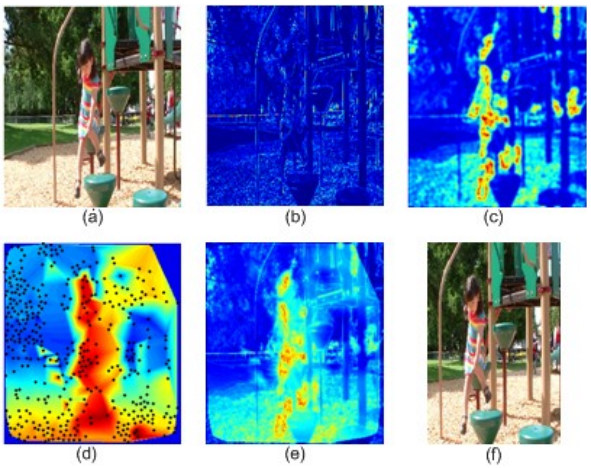

Figure 4. A frame from the sequence "Playground." In this example, both the global texture map and the saliency map identify the little girl as the most salient region. The combined map informs our method where to place the bounding box.

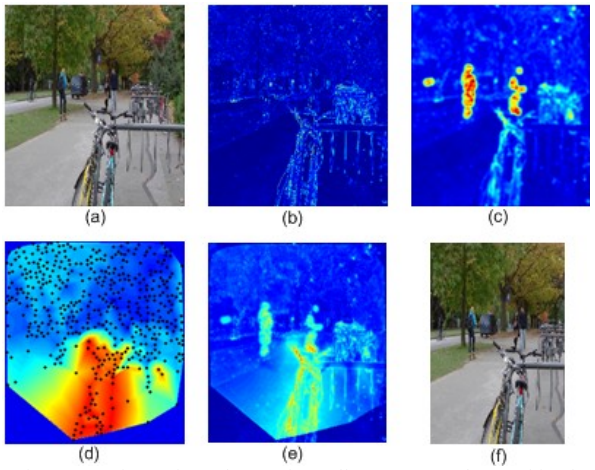

Figure 5. A frame from the "Main Mall" sequence. The combined map includes all the salient points in the frame and the bounding box is chosen accordingly.

6. REFERENCES

[1] Randy Conrod, "Demystifying Active Format Description," Harris Assured Communications, Mason, OH, USA, White Paper 2008.

[2] L. Itti, C. Koch, and 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, November 1998.

[3] C. Chamaret, J.C. Chevet, and O. Le Meur, "Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies," in IEEE International Conference on Image Processing (ICIP), Hong Kong, 2010, pp. 1077 - 1080.

[4] Dirk Walther and Christof Koch, "Modeling attention to salient proto-objects," Neural Networks, vol. 19, no. 9, pp. 1395-1407, November 2006.

[5] Christel Chamaret and Oliviere Le Meur, "Attention-based video reframing: validation using-eye-tracking," in 19th International Conference on Pattern Recognition, ICPR, Tampa, FL, USA, 2008, pp. 1-4.

[6] O. Le Meur, D. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," Vision Research, vol. 47, no. 19, pp. 2483-2498, September 2007.

[7] O. Le Meur, D. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pp. 802-817, May 2006.

[8] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in Int. Conf. Pattern Recognition, Barcelona, 2000, pp. 375-378.

[9] A. Maki, J. O. Eklundh, and P Norlund, "A computational model of depth-based attention," in International Conference on Pattern Recognition, vol. 4, Vienna, 1996, pp. 734-739.

[10] Yun Zhang, Gangyi Jiang, Mei Yu, and Ken Chen, "Stereoscopic Visual Attention Model for 3D Video," in Advances in Multimedia Modeling, Susanne Boll et al., Eds.: Springer Berlin / Heidelberg, 2010, pp. 314 - 324.

[11] David G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, November 2004.

[12] Manuel and Hornung, Alexander and Wang, Oliver and Poulakos, Steven and Smolic, Aljoscha and Gross, Markus Lang, "Nonlinear disparity mapping for stereoscopic 3D," ACM Transaction son Graphics, vol. 29, no. 4, pp. 75:1-75:10, July 2010.

[13] A Treisman and G Gelade, "A Feature-Integration Theory of Attention," Cognitive Psychology, vol. 12, no. 1, pp. 97-136, 1980.

[14] Quan Huynh-Thu, Marcus Barkowsky, and Patrick Le Callet, "The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives," IEEE transactions on broadcasting, vol. 57, no. 2, pp. 421-431, June 2011.

[15] Chenlei Guo, Qi Ma, and Liming Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion Fourier transform," in IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, 2008, pp. 1-8.

[16] Todd A. Ell and Stephen J. Sangwine, "Hypercomplex Fourier Transforms of Color Images," IEEE Transactions in Image Processing, vol. 16, no. 1, pp. 22-35, January 2007.

[17] Xuemei Zhang, Brian Wandell Stephen Engel, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," Nature (London), vol. 388, no. 6637, pp. 68-71, July 1997.

[18] N. Le Bihan S. Sangwine. (2005) Source Forge. [Online]. http://qtfm.sourceforge.net