This is a preliminary version of an article published by
B. Guthier, J. Kiess, S. Kopf, W. Effelsberg:
Seam Carving for Stereoscopic Video.
Proc. of IEEE Workshop on 3D Image/Video Technologies and Applications (IVMSP), June 2013.

SEAM CARVING FOR STEREOSCOPIC VIDEO

Benjamin Guthier, Johannes Kiess, Stephan Kopf, Wolfgang Effelsberg

Departement of Computer Science IV
University of Mannheim, Germany
{guthier, kiess, kopf, effelsberg}@informatik.uni-mannheim.de

ABSTRACT

In this paper, we present a novel technique for seam carving of stereoscopic video. It removes seams of pixels in areas that are most likely not noticed by the viewer. When applying seam carving to stereoscopic video rather than monoscopic still images, new challenges arise. The detected seams must be consistent between the left and the right view, so that no depth information is destroyed. When removing seams in two consecutive frames, temporal consistency between the removed seams must be established to avoid flicker in the resulting video. By making certain assumptions, the available depth information can be harnessed to improve the quality achieved by seam carving. Assuming that closer pixels are more important, the algorithm can focus on removing distant pixels first. Furthermore, we assume that coherent pixels belonging to the same object have similar depth. By avoiding to cut through edges in the depth map, we can thus avoid cutting through object boundaries.

Index Terms— Stereoscopic videos, 3D, seam carving, resizing

1. INTRODUCTION

Stereoscopic videos are becoming increasingly popular with more and more stereoscopic devices coming to the consumer market. Examples for these devices include TV screens, portable gaming consoles, smartphones, and video cameras. With the diversity of the available devices also comes the problem that the stereoscopic content does not fit all displays equally as it has a fixed resolution and aspect ratio. Therefore, the videos have to be adapted to fit the different screens. This process is called retargeting or resizing and is a research area that is well explored for 2D images and video.

This is not the case for stereoscopic content. While there are algorithms for the automatic resizing of stereoscopic images [1], to our knowledge there are no approaches for video yet that go beyond cropping the borders or linear scaling.

In this paper, we propose a content-aware algorithm for the automatic resizing of stereoscopic video based on *seam carving* [2]. For our method, we assume that the left and the right view of a video are given. The disparity map – the mapping between pixels in the left and the right frame – is calculated using existing algorithms [3]. In our approach, seams are searched in the left view and the disparity map simultaneously to preserve the depth information as well as possible. For temporal consistency, the seams from the previous frame are used as a reference for searching seams in the current frame. An extended version of this paper has been published as a technical report [4]. It contains details that were left out of this paper for lack of space.

The seam carving method for stereoscopic video presented in this paper focuses on the following:

- Consistency between the seams in the left and the right frame to preserve depth information.
- Temporal consistency between the seams in two consecutive video frames to avoid flicker.
- Use of depth information to preserve closer objects and to prevent cutting through object boundaries.

The outline of this paper is as follows: Section 2 presents the current state of the art of 2D video retargeting and stereoscopic image resizing. Our algorithm is described in detail in Section 3. Its achieved quality is evaluated in Section 4. Section 5 concludes the paper.

2. RELATED WORK

Retargeting or resizing describes the process of adapting an image or video to a different display resolution or aspect ratio. This process is well explored for 2D media and is a hot topic for stereoscopic media. *Seam carving* is one of the most prominent techniques and has been picked up by a lot of other researchers [1, 2, 5, 6, 7, 8]. To our knowledge, there is currently no work published on the resizing of stereoscopic video that goes beyond cropping the borders or uniform scaling.

Seam Carving is a technique for the content-aware retargeting of images and was first introduced by Avidan and Shamir [2]. A *seam* is a connected path of pixels from top to bottom or left to right. An energy function is used to evaluate the importance of each pixel in the image. The optimal seam which contains the pixels with the lowest overall energy is then detected and removed to reduce the size of the image by one column or row.

Rubinstein et al. extended the seam carving approach to the resizing of video [5]. They represent a video as a 3D spatial/time video cube. Graph cuts are used to find minimal energy seams. Also, they introduce *forward energy* which measures the energy that will be inserted by removing a seam rather than the energy that is deleted with the seam. Forward energy is used in many other seam carving approaches.

In the approach by Grundmann et al., seams are allowed to be spatially and temporally disconnected [6]. This gives the seams more flexibility and enables them to avoid crossing important objects. Additionally, a temporal coherence measure is proposed that allows a frame-by-frame computation with only the previous frame needed as reference.

Most recently, another seam carving algorithm for reducing the width of stereo *images* was presented by Basha et al. [1]. It jointly uses the information provided by both views for the computation of the energy map. The views are mapped onto each other by the disparity map. Their energy function considers forward energy [5] in both images as well as 3D energy. The latter is composed of forward energy in the disparity map, energy computed from depth and the confidence of the disparity estimation.

3. SEAM CARVING FOR STEREO VIDEO

The input to our algorithm is a video sequence consisting of left frames $I_t^L(x, y)$ and right frames $I_t^R(x, y)$. Sub- and superscripts are omitted whenever they are not required for understanding. Each frame of the input sequence is of size $w \times h$. The process of reducing their width starts by computing a disparity map between I^L and I^R to establish pixel correspondence among the views. We use semi-global block matching to compute the disparity map [3]. An energy function is then computed for the current frame that incorporates knowledge from both views at once. It judges each pixel's importance in the image. The energy values are accumulated row by row to calculate an accumulated energy map. Based on this map, seams of pixels with low energy are detected and removed from the two views. In the last step, the seam is also removed from the disparity map and disparity values are updated. The entire process is then repeated until the target width is reached. For lack of space, the description of our algorithm is kept short. Refer to the technical report for the details [4].

Our approach is focused on finding and removing vertical seams in a stereo pair. A vertical seam consists of exactly one x coordinate for each row in an image. It is a function of y. Removing a seam means deleting the seam pixel in each row and shifting all pixels to the right of the seam left by one. This reduces the width of the image by one. We distinguish between seams in the left and the right view by using the superscripts L and R. The pair of seams is connected by the disparity map D in the following way:

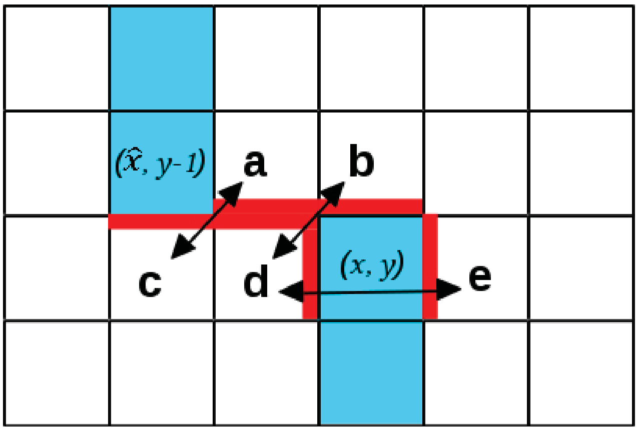

Fig. 1. The blue squares are pixels belonging to a seam. After removing it, the pixels labeled a through e change their neighbors. The affected sides of the pixels are marked in red. In this example, the forward energy is $|d-e|+|a-c|+|b-d|$.

$$S_i^R(y) = S_i^L(y) - D(S_i^L(y), y) \qquad (1)$$

3.1. Energy Function

The energy value of a pixel denotes its importance in the image. In our approach, the energy function is composed of appearance energy E_{app}, disparity energy E_{3D}, and temporal energy E_{temp}. Appearance energy measures edges in the intensity image that are introduced when removing a pixel. Disparity energy takes into account the removal of seams in the disparity map, as well as the depth of a pixel. Temporal energy helps to achieve temporal consistency by giving a higher energy to pixels that are far away from the seams of the previous frame. These three components are summed up to a total energy E:

$$E(x, y, \hat{x}) = \alpha_1 E_{app}(x, y, \hat{x}) + \alpha_2 E_{3D}(x, y, \hat{x}) + \alpha_3 E_{temp}(x, y)$$

Total energy is a function in three variables: x and y coordinate of the pixel and the horizontal location $\hat{x}$ of the seam pixel in the row above. Throughout this Section, the hat over a symbol is used when referring to values in the previous row or previous frame. See the technical report for our choice of weights α [4].

When removing seams from the left and right frames, pixels that were originally separated by the seam may become adjacent. This may introduce noticeable edges into the frames, which is generally undesirable. The effect of introducing new edges into the frames by removing seams is measured by appearance energy [5]. The appearance energy $E_{app}(x, y, \hat{x})$ at a pixel position (x, y) depends not only on the pixel position itself, but also on the horizontal position $\hat{x}$ of a potential seam pixel in the row above $(\hat{x}, y-1)$. This is illustrated in Figure 1. Appearance energy is composed of two parts:

$$E_{app}(x,y,\hat{x}) = E_{hor}(x,y) + E_{ver}(x,y,\hat{x})$$

They are horizontal (E_{hor}) and vertical energy (E_{ver}). When a pixel at (x,y) is removed, its left and right neighbors become adjacent, introducing a new edge. This is measured by horizontal energy which is simply the difference between the intensities of the left and the right neighbor:

$$E_{hor}(x,y) = |I(x-1,y) - I(x+1,y)|$$

If $x \neq \hat{x}$, removing a seam causes a shift between rows $y-1$ and y over the length of $|x-\hat{x}|$ (see Figure 1). This is measured by vertical energy:

$$E_{ver}(x,y,\hat{x}) = \begin{cases} \sum\limits_{k=\hat{x}+1}^{x} |I(k,y-1) - I(k-1,y)| & \text{if } \hat{x} < x \\ \sum\limits_{k=x+1}^{\hat{x}} |I(k-1,y-1) - I(k,y)| & \text{if } \hat{x} > x \end{cases}$$

The horizontal pixel positions x and $\hat{x}$ are mapped into the right frame by subtracting the disparity. Like this, the appearance energy is calculated for the left and the right view simultaneously. The final value for $E_{app}(x,y,\hat{x})$ is then obtained by adding the energy values of the two corresponding left and right pixels.

Disparity energy E_{3D} is composed of forward energy in the disparity map E_{disp}, the distance of a pixel from the camera E_{dist}, and the confidence in the disparity estimation E_{conf}. Our definition of disparity energy is similar to the one in [1]:

$$E_{3D}(x,y,\hat{x}) = E_{disp}(x,y,\hat{x}) + \alpha_4 E_{dist}(x,y) + \alpha_5 E_{conf}(x,y)$$

$E_{disp}(x,y,\hat{x})$ is defined in the same way as E_{app} above, except that it is computed over the disparity map instead of the intensity image. It prevents seams from introducing depth edges into the disparity map. The energy from object distance is simply defined as normalized disparity: $E_{dist} = D$. This makes closer objects less likely to be removed by seam carving, because they have a higher disparity and thus higher energy. In order to cope with noisy disparity measurements, we include E_{conf} into the disparity energy, which represents the confidence in the disparity measurement at a pixel. For a good disparity value, the color values of two corresponding pixels in the left and right frame should only differ by a small amount. E_{conf} is thus simply set to this color difference.

When applying seam carving frame by frame to a video, the seams take a different path in every frame. This introduces artificial motion into the frame which is perceived as a disturbing flicker artifact. To avoid flicker, it is necessary to make sure that seams do not differ from the seams in the previous frame by too much. This is done by adding temporal energy to the energy function as was shown in [6]. During the detection of the i-th seam in the *current* frame, the temporal energy E_{temp} for a pixel measures by how much the result differs if this pixel is removed instead of removing the i-th seam of the *previous* frame again.

More formally, when computing the i-th seam $S_i^L(y)$ in the left frame at time t, the i-th seam in the left frame at time $t-1$ is taken into account. This seam in the previous frame is denoted by $\hat{S}_i^L(y)$. If the exact same seam $\hat{S}_i^L(y)$ was used again as the i-th seam of the current left frame I_t^L, the resulting frame after removing the seam would be $\hat{I}_t^L$. Row y of frames I_t^L and $\hat{I}_t^L$ are shown on the right side of Figure 2. Frame $\hat{I}_t^L$ would have perfect temporal consistency, because the same pixels as in the previous frame were removed. For each pixel position (x,y) in the left frame, the temporal energy $E_{temp}^L(x,y)$ is thus computed as the difference between frame I_t^L as if it were carved by a seam going through pixel (x,y) and the perfectly consistent frame $\hat{I}_t^L$. Removing a seam pixel at position (x,y) in frame I_t^L means that all pixels to the right of x are shifted left by one. Hence, E_{temp}^L is defined as:

$$\begin{aligned} E_{temp}^L(x,y) &= \sum_{k=0}^{x-1} |I_t^L(k,y) - \hat{I}_t^L(k,y)| \\ &+ \sum_{k=x+1}^{w-i+1} |I_t^L(k,y) - \hat{I}_t^L(k-1,y)| \end{aligned}$$

E_{temp}^R is computed analogously by mapping x into the right frame. Total temporal energy E_{temp} is then obtained by adding the values of both views.

3.2. Finding and Removing Seams

After fully defining the energy function, it can be used to detect and remove seams with low energy in the video frames. Note that only one seam pair is detected and removed at a time, so the seam index i can be omitted.

In order to compute a pair of seams $S^L(y)$ and $S^R(y)$, the energy function is accumulated over each row of the frame, starting from the top. The result is an accumulated energy map $M(x,y)$. For each pixel position (x,y), all potential predecessor pixels $(\hat{x}, y-1)$ in the row above are considered. The predecessor leading to the lowest energy is chosen and the total energy is accumulated:

$$M(x,y) = \min_{\hat{x}} M(\hat{x}, y-1) + E(x,y,\hat{x})$$

The last row of the accumulated energy map $M(x,h-1)$ then contains the accumulated energy of a left seam ending in location $(x,h-1)$. The minimum of the entire last row marks the endpoint of a left seam with the lowest energy. It can be traced back to the first row by following the stored predecessors. Like this, the entire seam is defined. By using the disparity map, the seam is mapped into the right frame to

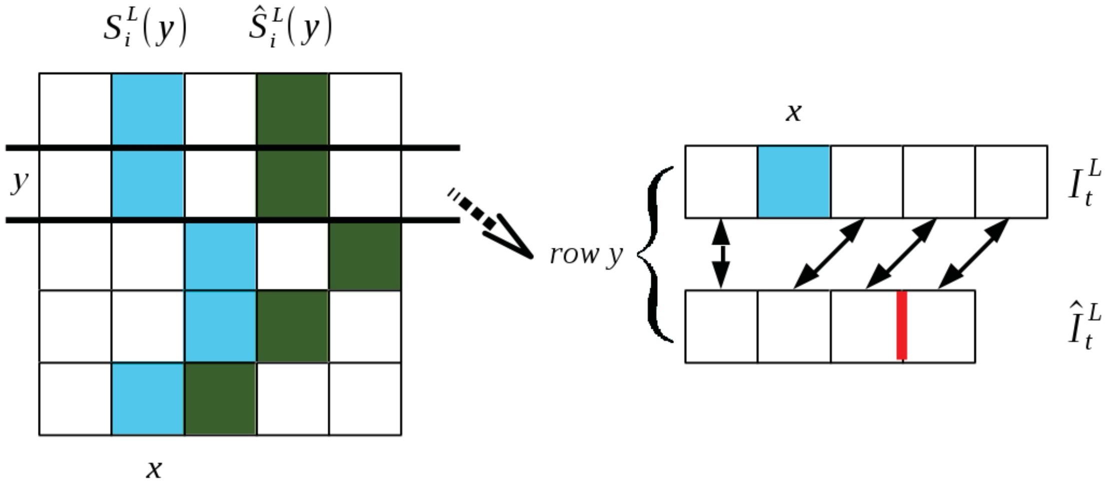

Fig. 2. The blue seam is a potential seam in the current frame. The green one is the unchanged seam $\hat{S}_i^L(y)$ from the previous frame. For pixel (x, y), temporal energy is computed as a sum of differences between the current frame I_t^L and the frame $\hat{I}_t^L$, which is the result of removing seam $\hat{S}_i^L(y)$ from I_t^L. The red line marks the pixel that was removed. Pairs of pixels for which the difference is calculated are marked with an arrow. The leftmost and rightmost pair of pixels have zero difference.

obtain the right seam (see Equation 1). The detected vertical seams $S_i^{L/R}$ for the left and the right view are now removed from their respective frame. All pixels to the right of a seam position $(S_i^{L/R}(y), y)$ are shifted left by one pixel.

For reasons of efficiency, the disparity map is not recomputed after the removal of each seam. Instead, the seam is also removed from the disparity map and the disparity values around the removed seam are updated [1]. Details on how this is done can be found in the technical report [4].

4. EVALUATION

We evaluated the achieved quality of our algorithm by resizing five challenging stereoscopic videos. The selected videos depict indoor and outdoor scenes with moving objects. As there is currently no other method for content-aware resizing of stereo videos, we compare our new technique to our implementation of [1]. It employs appearance and disparity energy and avoids removing occluded or occluding pixels. However, the energy function in [1] has no temporal component as it is a still image approach. In the following, we refer to our own approach as SV for "stereo video" and abbreviate the other method by SF for "stereo frame-wise".

The evaluation was a no-reference comparison where the test subjects only got to see the retargeted results, but not the original sequence. This is comparable to the real-world-situation where users only see the resized video on their devices. As test sequences, five stereo videos depicting indoor and outdoor scenes with moving objects were used. We refer to them as: ”dialog”, ”office”, ”street”, ”table” and ”walking”. Example frames of the resized sequences can be seen in 3. The full videos with a side-by-side frame format can be found online[1]. The original size of the videos was 480×270. They were resized to a size of 384×270, which is a reduction in width by 20%.

A total of 17 participants took part in the evaluation, three of which were knowledgeable in the field of video processing. For each video sequence, the results of the two algorithms were shown to the subjects in random order. The participants were first asked which of the two videos they prefer. Then the subjects assigned scores to the two sequences in four categories: deformation, cut-off objects, flicker, and distortion of the 3D effect. Each video in each category could be given a score of 1 (not noticeable), 2 (noticeable, but not disturbing), or 3 (noticeable and disturbing).

The evaluation showed that results of our stereo video approach were significantly preferred over the frame-wise approach without a temporal component. When asked which of the two compared videos has higher overall quality, the subject chose the video produced by our method 92% of the time. The complete table of scores given in the four categories are contained in the technical report [4].

The viewers' preference is mainly influenced by the improved temporal stability of our approach, which leads to considerably less flicker. Scores in the other three categories were largely the same for both approaches, as was to be expected. Deformations were noticed in both approaches equally but were classified as not disturbing (average scores SF: 2.13, SV: 1.82). Our algorithm performed slightly worse in the category of cut-off objects (SF: 1.34, SV: 1.52). Both scores are in the

[1] `http://ls.wim.uni-mannheim.de/de/pi4/research/projects/retargeting/`

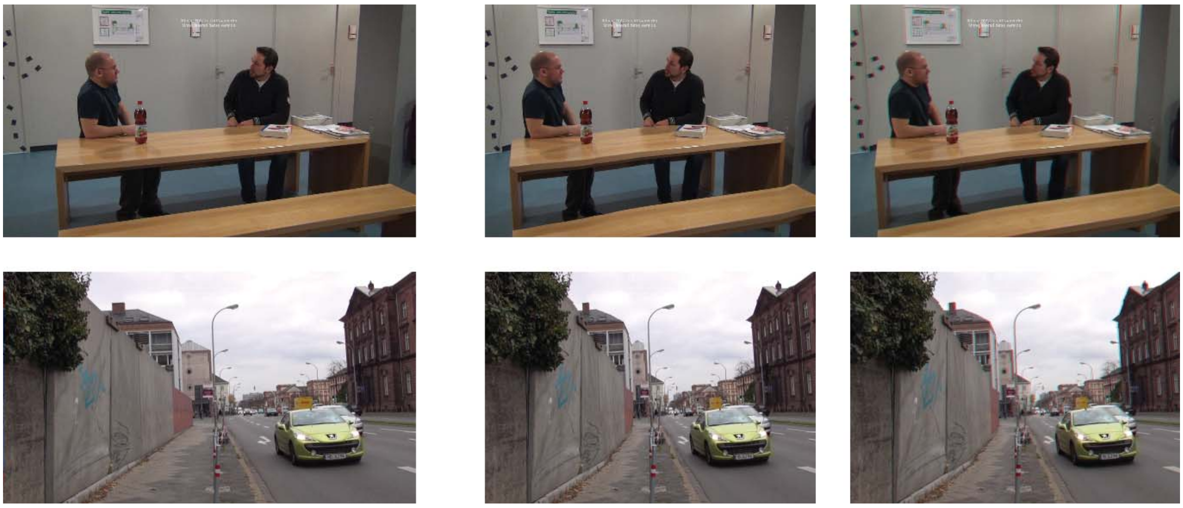

Fig. 3. Example frames from the test sequences ”table”, and ”street” that were used in our evaluation. The width of the videos was reduced by 20%. Left: left view of the original frame. Middle: left view of the resulting frame. Right: anaglyph (red/cyan) version of the resulting frame.

range that indicates that this artifact remained mostly unnoticed and never disturbing. Flicker is an artifact which nearly all participants found to be very disturbing in the videos that were resized using the SF approach. It received the worst possible score in almost all of the ratings in this category. Flicker was not noticed in the SV sequences most of the time (SF: 2.87, SV: 1.41). The 3D impression of the sequences achieved high scores in both approaches (SF: 1.39, SV: 1.25). The subjects did not notice an impairment of the 3D effect in the average.

5. CONCLUSIONS

We presented a seam carving technique for stereoscopic video. Our technique takes forward energy in the left and right view as well as the disparity map into account. Additionally, it calculates energy from depth and adds temporal consistency to the seams. Our evaluation showed that temporal consistency is an important criterion when applying stereo seam carving to video. Its absence leads to flicker and strongly decreases the perceived video quality.

Subjectively, the 3D effect was not impaired by seam carving. We believe that this effect may be too subtle to notice in a complex video scene. We did not find it necessary to detect and avoid occluded and occluding pixels in our approach. It was found that special treatment of such pixels has a negative effect on quality when the disparity map contains errors.

6. REFERENCES

[1] T. Basha, Y. Moses, and S. Avidan, “Geometrically consistent stereo seam carving,” in *ICCV*, nov. 2011, pp. 1816–1823.

[2] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” *ACM Trans. on Graphics*, vol. 26, no. 3, 2007.

[3] H. Hirschmüller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” in *In Proc. of the CVRP*, 2005, pp. 807–814.

[4] B. Guthier, J. Kiess, S. Kopf, and W. Effelsberg, “Stereoscopic seam carving with temporal consistency,” Tech. Rep. TR-2013-002, University of Mannheim, April 2013.

[5] M. Rubinstein, S. Avidan, and A. Shamir, “Improved seam carving for video retargeting,” *ACM Trans. on Graphics*, vol. 27, no. 3, 2008.

[6] M. Grundmann, V. Kwatra, Mei Han, and I. Essa, “Discontinuous seam-carving for video retargeting,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 569 –576.

[7] S. Kopf, J. Kiess, H. Lemelson, and W. Effelsberg, “FSCAV: Fast seam carving for size adaptation of videos,” in *Proc. of the Int. Conf. on Multimedia (MM)*, 2009, pp. 321–330.

[8] J. Kiess, B. Guthier, S. Kopf, and W. Effelsberg, “SeamCrop: Changing the size and aspect ratio of videos,” in *Proc. of the 4th Workshop on Mobile Video*, 2012, MoVid ’12, pp. 13–18.