SALIENCY-BASED FEATURE SELECTION STRATEGY IN STEREOSCOPIC PANORAMIC VIDEO GENERATION

Haoyu Wang[⋆]*, Daniel J. Sandin*[†]*, Dan Schonfeld*[⋆]

[⋆] Department of Electrical and Computer Engineering, University of Illinois at Chicago
851 S. Morgan Street, Chicago, IL 60607, U.S.A.
[†] Department of Computer Science, University of Illinois at Chicago
851 S. Morgan Street, Chicago, IL 60607, U.S.A.

ABSTRACT

In this paper, we present one saliency-based feature selection and tracking strategy in the feature-based stereoscopic panoramic video generation system. Many existing stereoscopic video composition approaches aim at producing high-quality panoramas from multiple input cameras; however, most of them directly operate image alignment on those originally detected features without any refinement or optimization. The standard global feature extraction threshold always fails to guarantee stitching correctness of all human interested regions. Thus, based on the originally commonly identified feature set, we incorporate the saliency map into the distribution of control points to remove the redundancy in texture-rich regions and ensure the adequacy of selected features in visual sensitive regions. Intuitively, under the guidance of saliency change in the video sequence, one grid-based feature updating strategy is operated between consecutive frames instead of the standard global feature updating. The experiments show that our method can improve the stitching quality of visual important region without impairment to the human less-interested regions in the generated stereoscopic panoramic video.

Index Terms— Stereoscopic Panoramic Video, Commonly-identified Feature, Saliency Map, Visual Sensitive

1. INTRODUCTION

Panoramic 3D video stitching is always a tough topic because of its challenges including temporal coherence, dominate foreground objects moving across views and camera jitters. Although many feature-based 3D panoramic video stitching methods have been proposed to solve those problems, most of them only focus on the improvement or optimization to fitted alignment parameters based on extracted features [1, 2, 3, 4]. The originally detected features from various algorithms [5, 6] are usually directly used for image alignment with equal importance and then be tracked in the video sequence with a global criterion for new feature detection operation. Although many complicated techniques and hardware-orientated solutions are proposed for high-quality stereo vision, they usually require expensive computation, complicated hardware set up and densely-sampled depth information [7, 8]. The goal of this study is to provide an efficient general feature-based strategy that could produce fewer artifacts or misalignment in those visual sensitive regions with sparsely sampled depth information.

In this paper, the proposed feature selection strategy is established on the construction of a commonly identified feature set at the first frame. Furthermore, we combine saliency and gradient map to represent the visual importance of each pixel and fairly distribute control points under the guidance of generated saliency energy map. Thereafter, one grid-based feature update strategy is employed to execute local feature detection and rejection instead of the conventional global feature update strategy in tracking stage.

2. RELATED WORKS

Video stitching is always a challenge task and becomes even more problematic when extended to stereo sense. Different feature-based stitching approaches have been proposed to generate high quality stereoscopic panoramic video. Jiang *et al.* proposed one algorithm for stitching multiple synchronized video streams into a single panoramic video with spatial-temporal content-preserving warping [1]. Li *et al.* proposed an wide-view video stitching method based on fast structure deformation [2]. Hamza *et al.* stabilized panoramic videos captured on portable platforms [3]. Perazzi *et al.* calculated optical flow to warp different views [4]. However, all these methods didn't consider any strategy to refine the original detected feature set and directly utilize them in image alignment and tracking stages. On the other hand, Guo *et al.* proposed a grid-based feature tracking method to produce more uniformly distributed features [9]. Zhu *et al.* selected corners based on the variance of region gray values to guarantees that corners are distributed proportionally to region texture information[10]. But both of them lack the scheme to ensure adequate features selection in all human interested regions.

Thus, inspired by the idea to minimize the saliency energy loss caused by the removal of image content in [11], we propose one feature selection strategy that could refine the originally detected feature set and redistribute control points according to the human visual attention. Furthermore, the local feature update strategy is also operated based on the temporal change of saliency energy in the later tracking stage intuitively.

3. PROPOSED FEATURE SELECTION AND TRACKING STRATEGY

3.1. Construction of Commonly-identified Feature Set

To provide a reliable matched feature set for our saliency-based selection, we firstly utilize the commonly-identified feature technique to describe common features from two pairs of input rectified stereoscopic images I_{L1}, I_{R1}, I_{L2}, and I_{R2} [12]. The score to evaluate the correspondence between four randomly choosen features d_1, d_2, d_3, d_4 is defined as:

$$\epsilon(d_1,d_2,d_3,d_4) = \sum_{i=1}^{3}\sum_{j=i+1}^{4} \|d_i.v - d_j.v\|^2 + \|d_1.y - d_3.y\|^2 + \|d_2.y - d_4.y\|^2 + \|\frac{f*b}{d_1.x - d_2.x} - \frac{f*b}{d_3.x - d_4.x}\|^2 \quad (1)$$

In each feature descriptor, vector $d_i.v$ stores gradient information and scalar pair $d_i.x, d_i.y$ represent the center point position of features. Generally, the evaluation score above contains six gradient difference terms, two vertical disparity terms, and one depth difference term. The symbol f is the focal length and b is the baseline.

Thus, the construction of the commonly identified feature set could be formulated as multiple optimization problems for each extracted feature descriptor. For each feature descriptor d_1 from the image I_{L1}, we can obtain the best-matched features in the other three images:

$$(\hat{d}_{1,2}, \hat{d}_{1,3}, \hat{d}_{1,4}) = \arg \min_{\substack{d_2 \in I_{L2} \\ d_3 \in I_{R1} \\ d_4 \in I_{R2}}} \epsilon(d_1,d_2,d_3,d_4) \quad (2)$$

Similarly, we can repeat above process for every feature descriptor in each image. According to the different sources of chosen feature descriptors for optimization, four images will produce four candidates of the commonly identified feature set: S_{L1}, S_{L2}, S_{R1}, and S_{R2}. Hence, the verified commonly identified feature set is set as the intersection of the above four candidates for uniqueness:

$$S_v = S_{L1} \cap S_{L2} \cap S_{R1} \cap S_{R2} \quad (3)$$

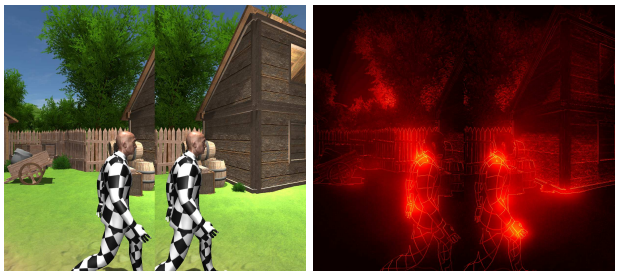


Fig. 1: Example of generated energy maps.

3.2. Saliency-based Feature Selection

Once we obtained the commonly-identified feature set for the four neighboring images at first frame, the next step is to select the reliable feature set for current frame stitching. Generally, the traditional global threshold often produces fewer features in poor texture areas because of the threshold will be biased by texture rich regions [13]. Thus, when some poor texture areas attract more attention from viewers, the inadequate number of control points will make most of them be considered as outliers during the RANSAC-based homograph estimation process. Hence, more misalignment or stitching errors are expected to appear in those regions. To improve the stitching quality of visual important areas, one saliency-based feature selection strategy is employed to adjust the selected control points into more reasonable distribution.

Our feature selection strategy starts with the generation of energy map, which indicates the visual importance of all pixel in each video frame. To generate visual sensitivity map with more sharp boundary, one energy fusion function [11] is used to combine the gradient map and GBVS-based saliency map [14, 15] as:

$$e(i,j) = \alpha_1 \cdot Gradient(i,j) + \alpha_2 \cdot GBVS(i,j) \quad (4)$$

where i,j represents the pixel of coordinate. Based on min-max normalization, the value of $Gradient(i,j)$ and $GBVS(i,j)$ are both normalized into [0,1]. Thus, for the four images in the first frame, we can compute the corresponding energy maps of their overlapping region: E_{L1}, E_{L2}, E_{R1} and E_{R2}. The generated energy maps of left view neighboring images is shown in Fig.1.

Then, we fragment each overlapping region into $M \times N$ grids: $\{G_{p,q}, p \in \{1,2,...M\}, q \in \{1,2,...N\}\}$. For each grid at pth row and qth column, the corresponding energy weight, $\hat{\omega}_{p,q}$, is defined as the normalized value of energy summation in the grid:

$$\omega_{p,q} = \sum_{(i,j)\in G_{p,q}} e(i,j) \quad (5)$$

$$\hat{\omega}_{p,q} = \frac{\omega_{p,q}}{\sum_{p,q}\omega_{p,q}} \quad (6)$$

The energy weight $\hat{\omega}_{p,q}$ represents the corresponding percentage of visual importance in the whole overlapping region. After running the above operations in the four regions:

E_{L1}, E_{L2}, E_{R1} and E_{R2}, we can use the average of four normalized energy weight as the commonly-identified weight of all four corresponding grids:

$$\omega^c_{p,q} = (\hat{\omega}^{L1}_{p,q} + \hat{\omega}^{L2}_{M,q} + \hat{\omega}^{R1}_{p,q} + \hat{\omega}^{R2}_{M,q})/4 \tag{7}$$

Given the total number of control points we are going to process, T, we can compute the number of features we need to select in each grid:

$$B_{p,q} = T \times \omega^c_{p,q} \tag{8}$$

Due to the over-sized features in texture rich grid, we need to remove some less-reliable or redudant feature squads based on our proposed ranking score. For each commonly-identified feature squad $\{d_{L1}, d_{L2}, d_{R1}, d_{R2}\}$, its ranking score consists of one matching confidence term and one disparity term:

$$\begin{aligned} R(d_{L1},d_{L2},d_{R1},d_{R2}) = \frac{\beta_1}{\epsilon(d_{L1},d_{L2},d_{R1},d_{R2})} + \\ \beta_2 \cdot [\|d_{L1}.x - d_{R1}.x\| + \|d_{L2}.x - d_{R2}.x\|] \end{aligned} \tag{9}$$

Since the small corresponding score between 4 control points from equation 1, $\epsilon(d_{L1}, d_{L2}, d_{R1}, d_{R2})$, implies high matching reliability of the feature squad, the matching confidence term is then defined as the reciprocal of it. The disparity term is incorporated to compensate those nearby objects that to any visible stitching errors.

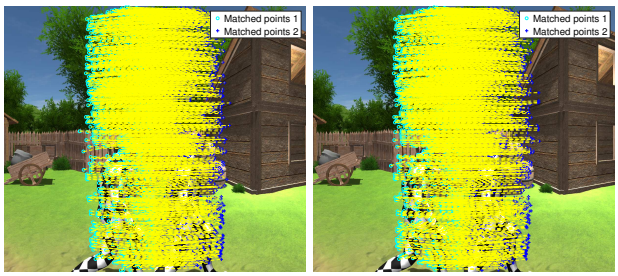
All commonly-identified feature squad in each grid are sorted in a descending order of our proposed ranking score R. The first $B_{p,q}$ commonly-identified feature squad in each grid are then regarded as the selected control points and will be tracked in consecutive frames. One example of matched feature squads before and after the saliency-based selection is depicted in Fig.2. It's appreantly that many reduandany control points are removed and those selected control point are distributed more uniformly.

3.3. Grid-based Local Feature Update Strategy

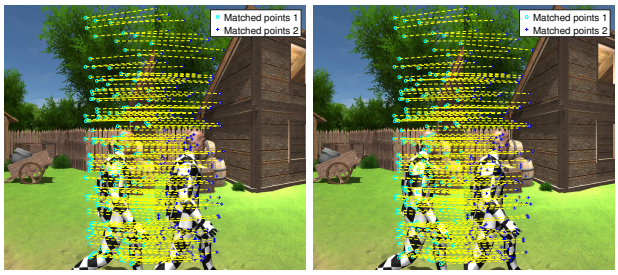
In conventional method, those detected features will be sent to various tracker which can estimate their position in later frames. The new feature detection will be operated only when the global number of tracked features drops below a given threshold. Unfortunately, the lost or mistakenly tracking of several key features may largely impair the stitching quality because of its failure to initialize the new feature detection operation. To avoid this kind of situation, we proposed the local saliency-based feature tracking strategy that focus on the temporal energy change of each grid. In other words, instead of stitching merely based on tracked features or newly-detected features, we update the features in each grid independently and operate image alignment based on these local hybrid feature set. The temporal energy change between previous frames t and current frame $t+1$ is defined as:

$$\delta_{p,q} = B^{t+1}_{p,q} - B^t_{p,q} \tag{10}$$

The positive $\delta_{p,q}$ means the new feature detection is needed in the grid of current frame, thus the local commonly-identified feature construction and matching ranking will be operated in grid $G_{p,q}$. The negative $\delta_{p,q}$ indicates some redundant control points need to be removed according to the ranking scores and the zero-valued $\delta_{p,q}$ implies all tracked features from the previous frame will be used for stitching of current frame.



(a) Matched commonly-identified features between neighbour cameras before saliency-based selection in left and rigth view.



(b) Matched commonly-identified features between neighbour cameras after saliency-based selection in left and rigth view.

Fig. 2: Comaprison of commony-identified feature squads for stitching before and after the saliency-based selection.

4. EXPERIMENTS

To demonstrate the improvement of our proposed feature selection and tracking strategy to stitching quality of the stereoscopic panoramic video, we implemented our saliency-based feature selection strategy (SFS) based on the framework of open-source panorama stitching software Hugin[16] and compared result with the standard method, as no feature selection strategy (NFS). The experiments data are synthesized outdoor scenes that describe one walking man in a circular path with different radius. In our experiments, the overlapping regions of four neighboring images are all divided into 10 by 5 grids. The α_1 and α_2 are set as 0.5. The two coefficients β_1 and β_2 in equation 9 are set as 0.7 and 0.3 .

4.1. Visual Improvement to Stereo Panoramic Video

Fig.3 shows several left view panoramas stitched by no feature selection strategy and our proposed strategy. In the areas marked by the blue rectangles in the top three panoramas, the walking man suffers from several visible stitching errors like the distortion of the head and the discontinues of the chest. In the top row of Fig.4, these evident monocular stitching

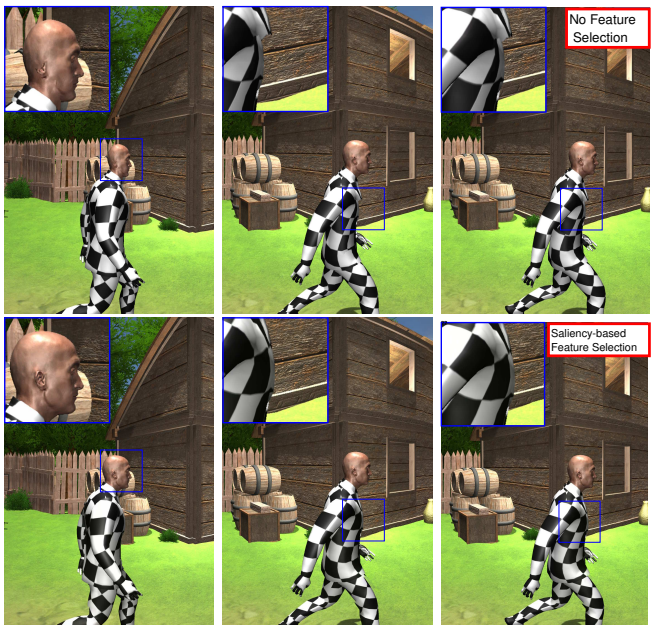


Fig. 3: Comparison of left view video stitching result between NFS and SFS.

errors deliver contradicted depth information of human head and chest in the stereoscopic video and result in serious viewing discomfort. This is because the feature detector fails to sample adequate control points in the suit with chess board texture. However, our proposed method can handle this problem and get the close walking man smoothly stitched. Based on the optimized distribution of control points that sample adequate percentage of features in those visual sensitive regions, the homography estimation will be operated under the guidance of the human attention. Thus, better monocular stitching quality and correctly embedded depth information in those visual sensitives region are expected.

4.2. Quantitative Comparison

To quantify the improvement of our proposed method to monocular frame stitching quality over standard selection strategy, the root means square error (RMSE) is used to evaluate the accuracy of alignment. For quantitative analysis of the stereoscopic panoramic video in the vertical direction, we measured the average vertical disparity of all matched features between left and right views. For the horizontal direction, we first consider the estimated depth from the original rectified image pair as the ground truth. The average distance of all matched features between the depth from stitched stereoscopic panoramas and the depth from the ground truth is then used as the metric to evaluate the performance of depth control. The numerical result of 20 frame synthetic outdoor scenes in different radius of circular path are shown in Table I. Each frame of stitched panoramaic video is scaled to 12000 by 3000 pixels for 360°x90°.

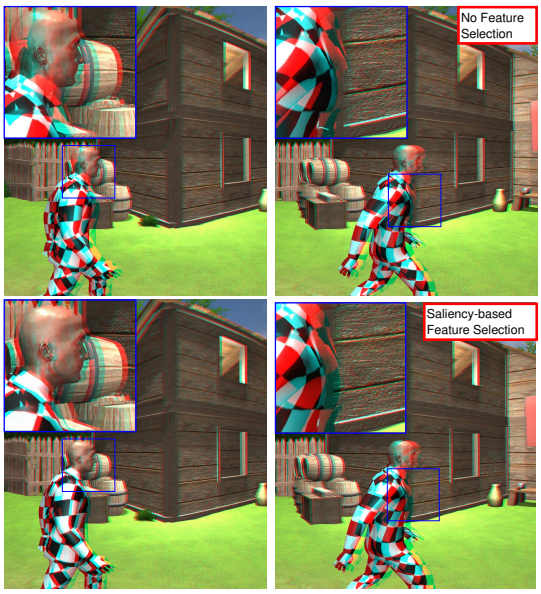


Fig. 4: Comparison of stereoscopic stitching result between NFS and SFS in red-cyan anaglyph version.

Table 1: Comparison result in circular path of different radius

	RMSE	Vertical Disp	Horizontal Dist
NFS+1.3m	$9.73px$	$0.19°$	$0.94°$
NFS+2.0m	$2.46px$	$0.13°$	$0.63°$
NFS+3.3m	$1.81px$	$0.10°$	$0.35°$
SFS+1.3m	$8.45px$	$0.05°$	$0.12°$
SFS+2.0m	$2.06px$	$0.06°$	$0.12°$
SFS+3.3m	$1.03px$	$0.05°$	$0.11°$

5. CONCLUSION

In this paper, we presented a feature selection and tracking strategy that optimizes the distribution of control points in panoramic video generation system. For this goal, we utilize the energy map that consists of saliency map and gradient map to compute the visual importance of each pixel. Based on them, we divide the overlapping region into grids and decompose the control points optimization problem into multiple ranking problems in each individual grid according to our proposed matching score. Afterwards, to maintain stitching correctness and temporal coherence, the control point set in each divided grid will be updated independently based on the change of energy. Some experiments based on synthesized data have been operated to prove the effectiveness of our method. In futrue works, more challenging tasks, such as mutiple movng objects, textureless moving object and complicated camera rigs setup, will be tested.

6. ACKNOWLEDGMENT

This publication is based on work supported in part by the National Science Foundation award CNS-1456638 for SENSEI.

7. REFERENCES

[1] Wei Jiang and Jinwei Gu, "Video stitching with spatial-temporal content-preserving warping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 42–48.

[2] Jing Li, Wei Xu, Jianguo Zhang, Maojun Zhang, Zhengming Wang, and Xuelong Li, "Efficient video stitching based on fast structure deformation," *IEEE transactions on cybernetics*, vol. 45, no. 12, pp. 2707–2719, 2015.

[3] Ameer Hamza, Rehan Hafiz, Muhammad M Khan, Yongju Cho, and Jihun Cha, "Stabilization of panoramic videos from mobile multi-camera platforms," *Image and Vision Computing*, vol. 37, pp. 20–30, 2015.

[4] Federico Perazzi, Alexander Sorkine-Hornung, Henning Zimmer, Peter Kaufmann, Oliver Wang, S Watson, and M Gross, "Panoramic video from unstructured camera arrays," in *Computer Graphics Forum*. Wiley Online Library, 2015, vol. 34, pp. 57–68.

[5] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[7] Kaimo Lin, Shuaicheng Liu, Loong-Fah Cheong, and Bing Zeng, "Seamless video stitching from hand-held camera inputs," in *Computer Graphics Forum*. Wiley Online Library, 2016, vol. 35, pp. 479–487.

[8] Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz, "Jump: virtual reality video," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 198, 2016.

[9] Heng Guo, Shuaicheng Liu, Tong He, Shuyuan Zhu, Bing Zeng, and Moncef Gabbouj, "Joint video stitching and stabilization from moving cameras," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5491–5503, 2016.

[10] Minchen Zhu, Weizhi Wang, Binghan Liu, and Jingshan Huang, "Efficient video panoramic image stitching based on an improved selection of harris corners and a multiple-constraint corner matching," *PloS one*, vol. 8, no. 12, pp. e81182, 2013.

[11] Tzu-Chieh Yen, Chia-Ming Tsai, and Chia-Wen Lin, "Maintaining temporal coherence in video retargeting using mosaic-guided scaling," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2339–2351, 2011.

[12] Haoyu Wang, Daniel Sandin, and Dan Schonfeld, "A common feature-based disparity control strategy in stereoscopic panorama generation," in *Visual Communications and Image Processing (VCIP), 2017 IEEE International Conference on*. IEEE, 2017.

[13] Matthias Grundmann, Vivek Kwatra, Daniel Castro, and Irfan Essa, "Calibration-free rolling shutter removal," in *Computational Photography (ICCP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–8.

[14] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[15] Xiaodi Hou, Jonathan Harel, and Christof Koch, "Image signature: Highlighting sparse salient regions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[16] "Hugin - panorama photo stitcher (version 2016.2.0)," .