3D Video Stabilization with Depth Estimation by CNN-based Optimization Supplementary Material

Yao-Chih Lee[1] Kuan-Wei Tseng[2] Yu-Ta Chen[2] Chien-Cheng Chen[2]
Chu-Song Chen[2] Yi-Ping Hung[2]
[1]Academia Sinica, Taiwan [2]National Taiwan University
[1]yclee1231@iis.sinica.edu.tw, [2]{kwtseng,r07922120,r09944015,chusong,hung}@csie.ntu.edu.tw

A. Appendix

A.1. Quantitative comparison on DeepStab [9]

We compare our Deep3D Stabilizer with state-of-the-art methods on DeepStab dataset [9]. DeepStab contains 61 pairs of synchronized unsteady and steady videos captured by two handheld cameras and one of which is mounted on a hardware stabilizer. We conduct the comparison on the test-split of DeepStab provided by Zhao *et al*. [10] in Table A.1. The stability of hardware stabilizer shows lower score than some software stabilizers may due to the limiting range of mechanical arm. Besides, we compute the cropping ratio with respect to input unsteady video to measure the view loss from input. Note that hardware stabilizer does not perform cropping process but we also compute the overlapping area with input video as reference. Our Deep3D shows the second best cropping score in the software stabilization methods. On the other hand, we measure the global distortion scores with respect to synchronized steady video since the unsteady input video may contain more undesired distortion than steady video (*e.g*., rolling shutter effects.) Our method has similar degree of global distortion to input videos since we do not handle distortion effects appeared in input video. Yet, StabNet [9] introduces even more distortion than input and ours.

Methods	Cropping w.r.t. Input	Distortion w.r.t. hardware	Stability
Input	-	**0.85**	14.2
Hardware	<u>0.84</u>	-	16.0
Adobe [4]	0.53	<u>0.84</u>	16.7
DIFRINT [1]	**0.98**	**0.85**	15.5
MeshFlow [5]	0.44	0.83	<u>17.4</u>
StabNet [9]	0.39	0.78	15.6
Ours	0.65	**0.85**	**17.5**

Table A.1. **Quantitative comparison on DeepStab [9].** The cropping ratio is measured with respect to input unstable video to measure the view loss from input video. Yet the global distortion score is measured with respect to synchronized steady video (*i.e*. hardware). The higher score of each metric indicates the better result. The best and the second best results are bolded and underlined, respectively.

A.2. Ablation studies on TUM RGB-D dataset [8]

We conduct the ablation studies on two random selected videos from TUM RGB-D dataset [8], which contains RGB-D videos and positioning groundtruth data captured by handheld Kinect and motion-capture system, respectively. Therefore, both the depth and camera pose estimation can be examined with the stabilized results. We also perform stabilization with groundtruth positioning and depth map as a reference 'GT'. The quantitative results of ablation on loss terms are shown in Table A.2. The different approaches show similar cropping and stability results yet the global distortion of the results are introduced due to the error of depth estimation. We admit that the errors of our 3D estimations are slightly large since we found a faithful 3D estimation would satisfy the quality of stabilization when replacing 2D with 3D transformation. Thus, instead of pursuing accuracy of depth and pose estimation, we reduce the computation time by parameter propagation for overlapping snippets and handling moving objects in post-processing.

Methods	3D estimation error↓		Video Stab.↑		
	depth	pose	C	D	S
Input	-	-	-	-	17.1
GT	-	-	0.51	0.90	20.4
w/o $\mathcal{L}_G$	1.60	1.04	0.52	0.46	20.0
w/o $\mathcal{L}_F$	0.97	0.54	**0.54**	0.74	**20.3**
Full	**0.74**	**0.33**	0.53	**0.84**	20.2

Table A.2. **Ablation studies on TUM RGB-D [8].** The error metrics of depth and pose are squared relative difference [2] and absolute trajectory error [8], respectively.

Category	Metric	Input	Liu *et al.* [3]	Deep3D (ours)
Regular	Cropping	-	0.44	0.79
	Distortion	-	0.85	0.97
	Stability	10.81	14.25	**15.30**
	# failure	-	3	0
Parallax	Cropping	-	0.74	0.76
	Distortion	-	0.85	0.92
	Stability	12.80	14.67	**14.74**
	# failure	-	5	0
Crowd	Cropping	-	0.56	0.70
	Distortion	-	0.89	0.93
	Stability	16.51	18.09	**18.37**
	# failure	-	7	0
Running	Cropping	-	0.46	0.49
	Distortion	-	0.77	0.89
	Stability	10.08	**15.49**	15.47
	# failure	-	3	0
Quick Rotation	Cropping	-	0.69	0.57
	Distortion	-	0.84	0.88
	Stability	18.86	19.23	**21.69**
	# failure	-	16	0
Zooming	Cropping	-	0.65	0.62
	Distortion	-	0.93	0.95
	Stability	16.96	19.42	**21.33**
	# failure	-	17	0

Table A.3. **Quantitative comparison with Liu *et al.* [3].** The scores are averaged over videos that are reconstructed and stabilized successfully. The higher score of cropping, global distortion and stability indicates the better result. The '# failure' stands for the number of videos failed in 3D reconstruction stage.

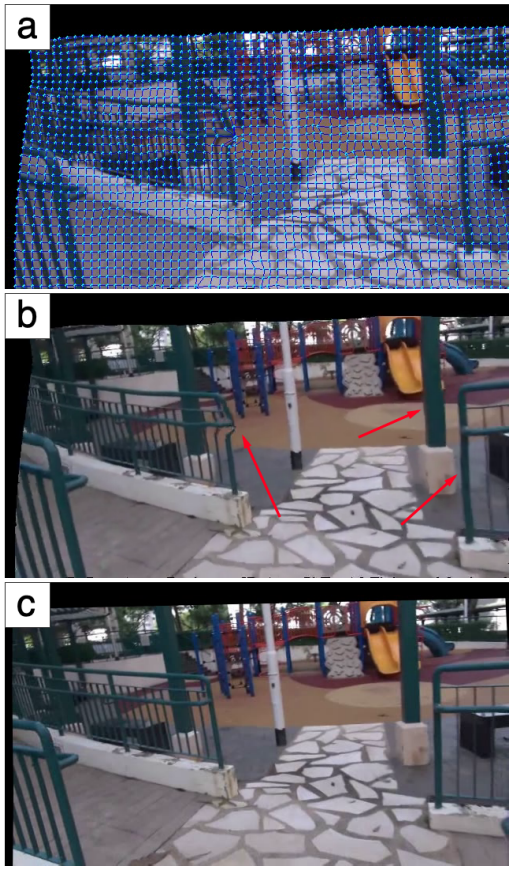

Figure A.1. **Visual comparison with Liu *et al.* [3].** (a) shows the optimized warping mesh of Liu *et al.* [3] guided by sparse structure points. (b) and (c) are the warped results of Liu *et al.* [3] and ours, respectively. The local distortion are pointed out by red arrows.

A.3. Comparisons with traditional 3D approach

We compare our Deep3D stabilizer with Liu *et al.* [3], which is a traditional 3D method using structure-from-motion to obtain the 3D path and sparse feature to guide the warping. We re-implement their content-preserving algorithm in Python and using COLMAP [7] for the 3D reconstruction stage. However, over one third of videos in NUS dataset [6] are failed to be stabilized since the tradiditonal SfM is often fragile. Therefore, we only compare the videos that are reconstructed successfully. As shown in Table A.3, our method outperforms Liu *et al.* [3] in general. Moreover, Liu *et al.* [3] often introduces local distortion due to some mis-matching structure points or dynamic objects (shown in Figure A.1.) On the other hand, [3] requires several hours in SfM stage for a short sequence and only reaches about 2 fps in warping optimization stage. In contrast, our optimization of 3D reconstruction and frame rectification only takes 641 ms and 29 ms per frame in average, respectively. In sum, our Deep3D stabilizer shows advantages of robustness of 3D reconstruction and computation efficiency.

References

[1] Jinsoo Choi and In So Kweon. Deep iterative frame interpolation for full-frame video stabilization. In *SIGGRAPH Asia*, 2019.

[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 3(January):2366–2374, 2014.

[3] Feng Liu, Michael Gleicher, Hailin Jin, and Aseem Agarwala. Content-preserving warps for 3d video stabilization. *ACM TOG*, 28(3):1–9, 2009.

[4] Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. Subspace video stabilization. *ACM TOG*, 30(1):1–10, 2011.

[5] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, and Bing Zeng. Meshflow: Minimum latency online video stabilization. In *ECCV*, pages 800–815. Springer, 2016.

[6] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM TOG*, 32(4):1–10, 2013.

[7] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[8] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, 2012.

[9] Miao Wang, Guo-Ye Yang, Jin-Kun Lin, Song-Hai Zhang, Ariel Shamir, Shao-Ping Lu, and Shi-Min Hu. Deep online video stabilization with multi-grid warping transformation learning. *IEEE TIP*, 28(5):2283–2292, 2019.

[10] Minda Zhao and Qiang Ling. Pwstablenet: Learning pixel-wise warping maps for video stabilization. *IEEE TIP*, 29:3582–3595, 2020.