(supplementary material)

# Reference Picture Synthesis for Video Sequences Captured with a Monocular Moving Camera

Hossein Bakhshi Golestani, Christian Rohlfing, and Jens-Rainer Ohm
Institute for Communications Engineering
RWTH Aachen University
Aachen, Germany.
{golestani, rohlfing, and ohm}@ient.rwth-aachen.de

This is the supplementary material for the paper submitted to VCIP 2019. In the following, first the detailed algorithm is given, then 3D reconstruction simulation results (e.g. point clouds, camera parameters, and 3D meshes) are depicted, and finally BD-PSNRs are reported.

## I. Detailed Algorithm

In video coding, Consider a 2-D video sequence captured by a monocular moving camera. The scene could be either static or dynamic. First, let's introduce notation. Input images $\mathrm{I} = \{\mathrm{I}_i \mid i = 1,2,\dots,N_{\mathrm{I}}\}$ are the pictures in the video sequence and $G$ is the GOP size. Key-frames (KFs), $\mathrm{I}_{\mathrm{KF}} = \{\mathrm{I}_i \mid i = G, 2G, \dots, \lfloor N_{\mathrm{I}}/G \rfloor\}$, are the first frame of each GOP. $M$ is intra period and $\mathrm{I}_{\mathrm{IRAP}} = \{\mathrm{I}_i \mid i = M, 2M, \dots, \lfloor N_{\mathrm{I}}/M \rfloor\}$ are Intra Random Access Points (IRAPs). $\mathrm{Mesh}_i$ is the 3-D mesh estimated from all KFs with indices less than or equal to $i$. $\mathrm{DM}_i$, $\mathrm{SP}_i$, and $\mathrm{CP}_i$ are the estimated Depth Map, Synthesized Prediction and Camera Parameters for $\mathrm{I}_i$. $\mathrm{R}_{\mathrm{Left}}$ and $\mathrm{R}_{\mathrm{Right}}$ are the left and right references for 3D warping.

```
Algorithm I
------------------------------------------------------------
Step1: Camera Parameters Estimation
- Apply SfM [1] to all raw frames.
- Save the parameters at the encoder side.
- Compress and send them to the decoder side.

# Encoding KFs ($i = nG$)

Step2: Depth map Generation
-   Apply Multi-View Stereo (MVS) [2] to all previous KFs
    ($i = 0, G, 2G, \dots, (n-1)G$) in order to estimate
    $\mathrm{Mesh}_{(n-1)G}$.
-   Back project (3D to 2D) $\mathrm{Mesh}_{(n-1)G}$ to previous KF
    ($\mathrm{I}_{(n-1)G}$) in order to generate $\mathrm{DM}_{(n-1)G}$.
Step3: 3D Warping
-   $\mathrm{R}_{\mathrm{Left}} = \mathrm{I}_{(n-1)G}$.
-   Apply bi-directional 3D warping to $\mathrm{R}_{\mathrm{Left}}$, and its
    corresponding depth maps to synthesize $\mathrm{SP}_i$.
-   No SP is provided for $\mathrm{I}_{\mathrm{IRAP}}$.


# Encoding BFs

Step2: Depth map Generation
-   Apply MVS to all KFs in order to estimate $\mathrm{Mesh}_{N_{\mathrm{I}}}$.
-   Back project (3D to 2D) $\mathrm{Mesh}_{N_{\mathrm{I}}}$ to all frames $\mathrm{I}_i$ with
    TID∈{1,2,3} in order to generate $\mathrm{DM}_i$.
Step3: 3D Warping
-   For each target B-Frame $\mathrm{I}_i$, find two references ($\mathrm{R}_{\mathrm{Left}}$ and
    $\mathrm{R}_{\mathrm{Right}}$):
    If $\mathrm{mod}(i,2) = 1$, $\mathrm{R}_{\mathrm{Left}} = \mathrm{I}_{i-1}$, $\mathrm{R}_{\mathrm{Right}} = \mathrm{I}_{i+1}$.
    If $\mathrm{mod}(i,G) = G/2$, $\mathrm{R}_{\mathrm{Left}} = \mathrm{I}_{i-G/2}$, $\mathrm{R}_{\mathrm{Right}} = \mathrm{I}_{i+G/2}$.
    If $\mathrm{mod}(i,G) = G/4$, $\mathrm{R}_{\mathrm{Left}} = \mathrm{I}_{i-G/4}$, $\mathrm{R}_{\mathrm{Right}} = \mathrm{I}_{i+G/4}$.
-   Apply bi-directional 3D warping to $\mathrm{R}_{\mathrm{Left}}$, $\mathrm{R}_{\mathrm{Right}}$ and their
    corresponding depth maps to synthesize $\mathrm{SP}_i$.
```

## II. 3D Reconstruction Simulation Results

In Table I, sample frames of the following tested sequences are shown.

1- Sintel [1], 4096×1744, 185 frames,

2- DayLightRoad, 3840×2160, 273 frames,

3- ParkRunning, 3840×2160, 281 frames,

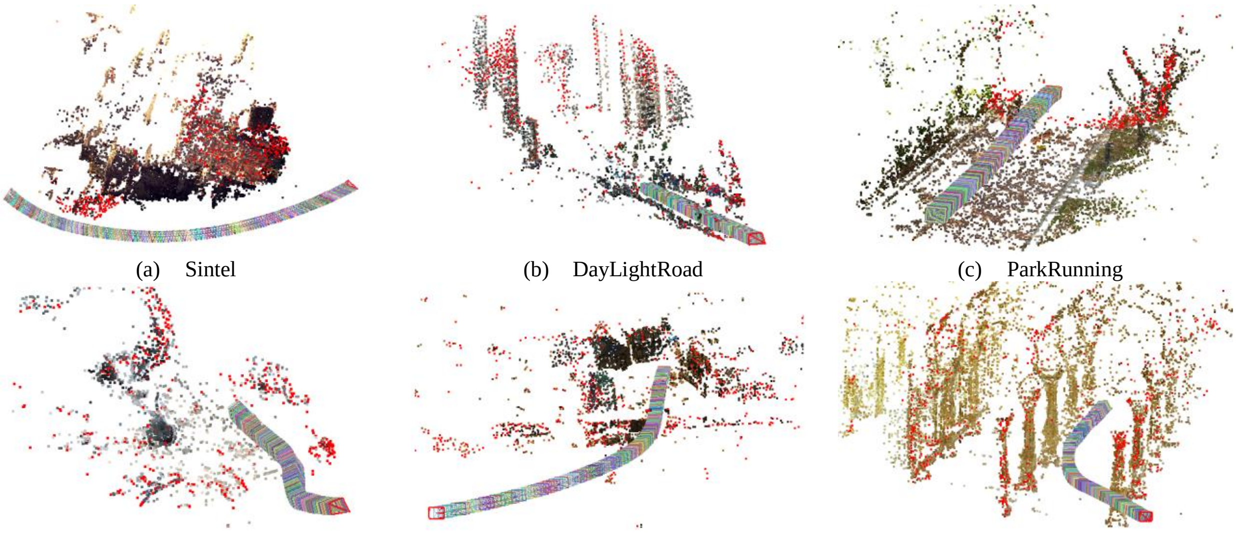
4- IceRock, 3840×2160, 281 frames,

5- GTFly, 1920×1088, 249 frames,

6- IndianBuilding [2], 1920×1080, 281 frames.

TABLE I. Sample Frames of the tested Sequences

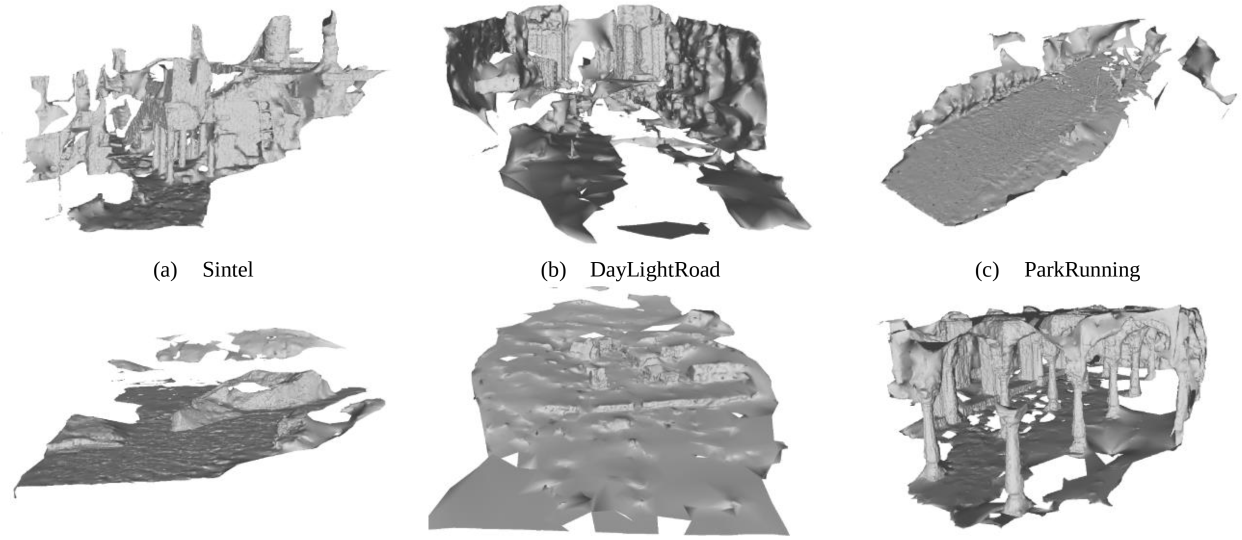
| Sequence | Sample frames | |
|---|---|---|
| Sintel | | |
| DayLightRoad | | |
| ParkRunning | | |
| IceRock | | |
| GTFly | | |
| IndianBuilding | | |



(a) Sintel (b) DayLightRoad (c) ParkRunning

(d) IceRock (e) GTFly (f) IndianBuilding

Fig. 1. The output of SFM (camera parameters and scene point cloud) for the tested sequenced



(a) Sintel (b) DayLightRoad (c) ParkRunning

(d) IceRock (e) GTFly (f) IndianBuilding

Fig. 2. The output of Multi-View Stereo (un-textured mesh) for the tested sequenced

As can be seen, the first three sequences contain moving objects, while the rest captured stationary scenes. The tested sequences were fed into SfM [3] and the outputs are reported in Fig. 1. The visual output of SfM, including sparse point cloud and estimated camera parameters. The accuracy of camera calibration could decrease if cameras with large baseline or highly compressed input images are used. Since SfM is applied to raw frames (key-frames and B-frames), usually one faces neither the large baseline problem nor having not enough corresponding points problem caused by low-quality input images. Fig. 2 shows the extracted 3D model from all key-frames with QP=29. 3D mesh has been reconstructed based on the minimum s-t cut solution, but with an emphasis to reconstruct weakly supported surfaces [4]. Note that we were not able to reconstruct very far away structures, for which it is very difficult to extract and match features accurately. Since the extracted features of moving objects are not consistent, extracting the 3D model of those objects is impossible. Also, homogeneous areas like sky could not be reconstructed because not enough features can be extracted in these areas.

## III. BD-PSNR

In order to save space, in the original paper only BD-Rates are presented. In this document, the BD-PSNR (dB) [5] values are reported (Table II). As can be seen, BD_PSNR behaves the same as BD-PNSR for different methods. In summary, the wider median filter performs better, the hierarchical method outperforms the KFs-Only scheme, and AR mode shows slightly better results compared to RR mode.

## References

[1] https://durian.blender.org/

[2] http://www.Free4kFootage.com/

[3] J. L. Schönberger and J. Frahm, "Structure-from-Motion Revisited," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113, 2016.

[4] M. Jancosek, T. Pajdla, “Multi-View Reconstruction Preserving Weakly-Supported Surfaces”, IEEE CVPR, pp. 3121-3128, 2011.

TABLE II. The Comparison of different methods in terms of BD-PSNR (dB) – Anchor: HEVC (HM16.7).

| Sequences | | [3] | | | Proposed Method | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Original | Modified Version | | AR Mode | | RR Mode | | RR Mode<br>+ Camera Parameters Overhead | |
| | | | | | KFs-Only | Hierarchical | KFs-Only | Hierarchical | KFs-Only | Hierarchical |
| | | w=3 | w=3 | w=15 | w=15 | w=15 | w=15 | w=15 | w=15 | w=15 |
| Class A | Sintel | +0.08 | +0.13 | +0.14 | +0.25 | +0.32 | +0.23 | +0.31 | +0.23 | +0.30 |
| | DayLightRoad | +0.04 | +0.07 | +0.07 | +0.09 | +0.11 | +0.08 | +0.10 | +0.08 | +0.10 |
| | ParkRunning | +0.08 | +0.12 | +0.13 | +0.16 | +0.18 | +0.15 | +0.18 | +0.15 | +0.18 |
| | Avg. Class A | +0.07 | +0.11 | +0.11 | +0.17 | +0.20 | +0.15 | +0.20 | +0.15 | +0.19 |
| Class B | IceRock | +0.32 | +0.33 | +0.37 | +0.48 | +0.51 | +0.47 | +0.50 | +0.46 | +0.50 |
| | GTFly | +0.19 | +0.25 | +0.25 | +0.45 | +0.48 | +0.44 | +0.47 | +0.44 | +0.46 |
| | IndianBuilding | - | +0.32 | +0.34 | +0.55 | +0.55 | +0.48 | +0.55 | +0.47 | +0.54 |
| | Avg. Class B | +0.25 | +0.30 | +0.32 | +0.49 | +0.51 | +0.46 | +0.51 | +0.46 | +0.50 |
| Average | | +0.14 | +0.20 | +0.22 | +0.33 | +0.36 | +0.31 | +0.35 | +0.30 | +0.34 |

[5] G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” ITU-T SG16/Q6VCE, Austin, USA, Tech. Rep. Doc. VCEG- M33, 2001.