Fast and fine disparity reconstruction for wide-baseline camera arrays with deep neural networks

T.Barrios, J. Gerhards, S. Prévost and C. Loscos

LICIIS laboratory, University of Reims Champagne-Ardenne, France

Abstract

Recently, disparity-based 3D reconstruction for stereo camera pairs and light field cameras have been greatly improved with the uprising of deep learning-based methods. However, only few of these approaches address wide-baseline camera arrays which require specific solutions. In this paper, we introduce a deep-learning based pipeline for multi-view disparity inference from images of a wide-baseline camera array. The network builds a low-resolution disparity map and retains the original resolution with an additional up scaling step. Our solution successfully answers to wide-baseline array configurations and infers disparity for full HD images at interactive times, while reducing quantification error compared to the state of the art.

CCS Concepts

• ***Computing methodologies*** → ***Computational photography; 3D imaging; Neural networks; Reconstruction;***

1. Introduction

Photogrammetric 3D reconstruction of a scene from a set of color images is needed by a wide range of applications, like the cultural heritage, autonomous driving, etc. With rectified, evenly spaced camera arrays and parallel optical centers, 3D reconstruction takes the form of *disparity* reconstruction, i.e., estimating an offset from each pixel of a view to an adjacent view of the camera array.

Deep learning methods have been widely explored and have proven to be well adapted to infer disparity from camera pairs [LJBB20] and light field cameras [HJKG16]. Much fewer solutions were proposed for wide-baseline camera arrays [LWZL21] which need to be addressed differently since the range of disparity values are in a different scale. In this paper, we propose a pipeline for multi-view disparity inference from wide-baseline RGB camera array using deep neural networks. An upscaling approach on the disparity maps, guided by input color images, is used to retain the original definition, processing FullHD images at interactive times.

2. Related work

While many camera configurations and targeted solutions have been proposed, we focus our review on aligned camera configuration, where the disparity concept applies. Several traditional reconstruction methods were proposed, using image correspondence with Markov Random Fields [Hua19], superpixels [CBG20] and view consistency-based refinement [CBG20] [SBV*17]. While they offer good quality results and, for more recent methods, fast computation time [CBG20] [SBV*17], they are often limited by design, making a fixed number of hypotheses and leading to disparity maps with discrete values and a quantification error in the output. Moreover, Li et al. [LWZL21] proposed an end-to-end deep learning solution for both short- and wide-baseline camera arrays. While increasing reconstruction accuracy this approach suffers from scalability issues on current GPUs.

3. Proposed method

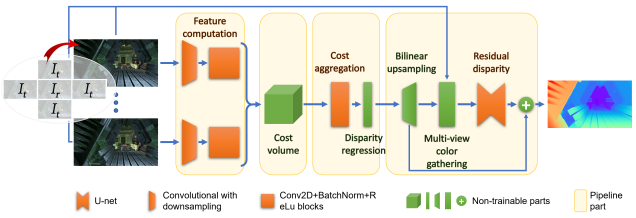

Figure 1: *Overview of our network: from the input (set of 4 RGB images $\{I_t\}$ and reference image (I_r) to the final disparity map.*

We propose a deep neural network that takes as input one reference color image I_r and a set of four color images $\{I_t\}$ in its neighborhood viewpoints of a camera array (sides, top and bottom). It outputs a single disparity map associated to the reference image. Similarly to [KFR*18], it first computes a low-resolution disparity map, and then, enhances its resolution to its full size. Our pipeline consists of four parts (see fig. 1).

Part 1 - Feature computation and downscaling: We use the feature computation network of [KFR*18] to downscale the image while increasing the number of features per pixel.

Part 2 - Cost volume computation: We compute cost features for

DOI: 10.2312/egp.20221007

each pixel and disparity candidate by subtracting for each the image features. Formally, for each pixel (i, j) on I_r and for each disparity candidate δ $in\{\delta_{min}/8, \delta_{min}/8+\delta step, .., \delta_{max}/8\}$, we compute a two-view cost-volume C between I_r and I_t from features f:

$$C(I_r, I_t, i, j, \delta) = f(I_r, i, j) - f(I_t, i+\delta_i, j+\delta_j) \quad (1)$$

with $\delta = \delta(I_r, I_t, \delta)$ the vertical/horizontal offset from I_r to I_t for disparity candidate δ. We can then define the global cost-volume C_v by concatenating the 4 two-view feature (one for each I_t).

Part 3 - Cost aggregation and disparity computation: This step aims at attributing a similarity score to each pixel and disparity candidate. Similarly to [KFR*18], we used a set of six 3D Convolution+BatchNormalization+ReLU. Each layer has an output of 64 channels. A final 3D Convolution layer, with no normalization nor activation with a single output channel, gives a final score $S_r(i, j, \delta)$ per pixel and disparity candidate. A first, downsampled disparity map is then computed by applying the *soft argmax* function to the scores. We then upscale it to input resolution disparity map Δ_r^u with bilinear upsampling before the refinement step.

Part 4 - Multi-view disparity refinement: We first fetch color data on target images $\{I_t\}$ corresponding to the target pixel of the disparity map Finally, we feed this image to a U-shaped 2D convolutional network, similar to the one used in [SS20] to compute residual disparity Δ_r^r. Then, we add it to the coarse disparity map in order to compute the final disparity map $\Delta_r = \Delta_r^u + \Delta_r^r$.

Training : We collected free-to-use 3D models and textures on various websites, which were randomly associated and positioned to create a 3D scene. From this scene, we take a set of 16 images from a 4×4 virtual camera array with disparity maps. We generated aorund 800 sets, each of FullHD resolution. Disparity values range from 0 to 270. For each set of the training dataset, 4 subsets of been created for training our network, using the four central views of our camera array as reference images. To avoid memory issues and to further increase variation in our dataset, iterations are performed on random crops of 960×540.The training loss is computed with both the coarse (Part 3) and fine (Part 4) output of the network.

4. Experiments

4.1. Method efficiency compared to state of the art

We tested our method on the *WLF hand designed test* dataset of [LWZL21]. We computed a disparity map only for the center view of the array. The camera array is 9×9, so we took the top-center, center-right, bottom-center and center-left input of the array as target images. δ_{min} and δ_{max} are set to 0 and 50 with δ_{step} .25.

The results of our method compared to existing methods [Hua19] [CBG20] [LWZL21] are shown in table 1 and figure 2. Results from [LWZL21] and [Hua19] are presented using the 9x9 camera grid while results from [CBG20] are presented using every other camera of the array because of memory issues. Our results show that our method outperforms the reconstruction quality of state-of-the-art methods in several aspects. First, the inference time is much smaller. Second, our reconstruction is higher in precision measured by the quantification error (bad 0.15).

Table 1: *Results for wide-baseline Light Field compared to state of the art. Computation times were measured on (*)CPU, (**)GPU.*

Method	Bad 0.15 (%)	Bad 0.3 (%)	Bad 0.6 (%)	Bad 1 (%)	Time (s)
[Hua19]	37.79	**5.32**	**2.97**	**2.55**	600*
[LWZL21]	15.04	7.05	3.95	2.80	40*
[CBG20]	25.71	10.67	4.15	3.22	1.6**
Ours	**14.09**	8.13	4.89	3.30	**0.5****

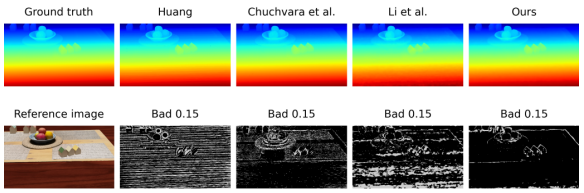

Figure 2: *Disparity recovery for a scene and quantification error (disparity error above 0.15). From left to right column: Ground truth, [Hua19], [LWZL21], [CBG20], ours.)*

5. Conclusion and future work

We presented a neural network, disparity-based 3D reconstruction method for wide-baseline camera arrays processing FullHD images at interactive times. We showed that our method reduces quantification error and inference time compared to the state of the art. As future work, we will adapt the method to generate a disparity map for each view while maintaining both speed and accuracy.

Acknowledgements. This work is funded by ANR-ReVeRY (ANR-17-CE23-0020) and used the URCA computing facilities of *Centre Image* and *ROMEO*.

References

[CBG20] CHUCHVARA A., BARSI A., GOTCHEV A.: Fast and accurate depth estimation from sparse light fields. *IEEE Transactions on Image Processing 29* (2020), 2492–2506. 1, 2

[HJKG16] HONAUER K., JOHANNSEN O., KONDERMANN D., GOLDLUECKE B.: A dataset and evaluation methodology for depth estimation on 4d light fields. Asian Conference on Computer Vision, Springer. 1

[Hua19] HUANG C.-T.: Empirical bayesian light-field stereo matching by robust pseudo random field modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41* (Mar 2019), 552–565. 1, 2

[KFR*18] KHAMIS S., FANELLO S., RHEMANN C., KOWDLE A., VALENTIN J., IZADI S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. *arXiv:1807.08865 [cs]* (Jul 2018). 1, 2

[LJBB20] LAGA H., JOSPIN L. V., BOUSSAID F., BENNAMOUN M.: A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. 1

[LWZL21] LI Y., WANG Q., ZHANG L., LAFRUIT G.: A lightweight depth estimation network for wide-baseline light fields. *IEEE Transactions on Image Processing 30* (2021), 2288–2300. 1, 2

[SBV*17] SABATER N., BOISSON G., VANDAME B., KERBIRIOU P., BABON F., HOG M., GENDROT R., LANGLOIS T., BURELLER O., SCHUBERT A., ET AL.: Dataset and pipeline for multi-view light-field video. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, p. 1743–1753. 1

[SS20] STUCKER C., SCHINDLER K.: Resdepth: Learned residual stereo reconstruction. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, p. 707–716. 2