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

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

Université de Reims Champagne-Ardenne, LICIIS, LRC DIGIT, France

Overview

Our goal is to reconstruct a 3D scene for wide-baseline arrays.

We propose a pipeline for **multi-view disparity inference from color images of a wide-baseline camera** array using deep neural networks , computing first a low-scale disparity map before being upscaled guided by input color images.

This pipeline allows us to reduce quantification error compared to state-of-the-art methods, and to process **FullHD** images at **interactive times**.

Problem Statement

- Photogrammetric 3D reconstruction of a scene from a set of color images
  ↳ Different approaches: varying on the type of input (number of images, camera configuration), the context and the objective
- **Context:**
  ↳ Wide-baseline camera array
  With a configuration of rectified, evenly spaced camera arrays, 3D reconstruction takes the form of disparity reconstruction.

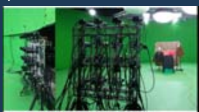

Fig 1: Camera example with a 20 cm baseline

  ↳ Exploitation of simplified epipolar geometry principle to compute disparity [1] (see fig 2)
- **Target :** δ in [0-250], baseline of 20 cm
- **Goal:** Multi-view depth estimation with less fine errors while maintaining a high resolution of the images and an interactive computation time.

Related Work

- Two categories of reconstruction methods
  - *Traditionnal methods* : using image correspondence [3] [5] superpixel [6] and multi-view consistency based refinement [5] [6]
    ↳ Good result but quantification error due to fixed number of hypotheses
  - *Deep-Learning* (DL) as [4]: increase reconstruction accuracy and reduce quantification error, but suffers from scalability issues on current GPUs

  ↳ Therefore, we chose to retain the DL categorie
- Testing Data: few datasets exist for multi-view camera arrays, especially with depth/disparity ground truth (GT), but we have:
  - Real, rectified FullHD data - Sabater *et al.* [5] with: δ in [0-300], 4x4 array, baseline of 7 cm
  - Virtual FullHD data with disparity GT - Li *et al.* [4] with: δ in [0-50], 9x9 virtual array.

  ↳ As [4] has too little data for our network and [5] has no disparity ground truth, we have composed our own dataset for training.

Acknowledgement

ANR-ReVeRy (ANR-17-CE23-0020)

Network Overview

Inputs :
- 1 Reference image : I_r
- Target images : { I_t }

Outputs :
- 1 Disparity map for I_r

Part 1 - Feature computation and downscaling

Goal: Increase per-pixel information and reduce image resolution to decreasing inference time.

Principle: Two sub-networks, see [7]:
1) Downscaling of 1/8 per dimension, with 3 sets of 2D-convolutional layers with a stride of 2.
2) Six 2D Convolution blocks = Convolution+BatchNormalization+LeakyReLU.

Output: 1/8 image with 32 channels for each image (I_r and I_t)

Part 2 - Cost volume computation

Goal: Feature computation for each pixel (i, j) and disparity candidate δ.

Principle: 1) Two-view cost-volume C , between I_r and I_t from features f with:

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

δ_i / δ_j : horizontal / vertical offset from reference image I_r to target image I_t for disparity candidate δ (see fig. 2).

2) Global cost-volume C_v as the concatenation of $C(I_r,I_t,i,j,\delta)$ set.

Output: $32.|\{I_t\}|$ channels, where N_t is the number of target images. for each pixel abd disparity candidate $\delta \in \{\delta_{min}/8, \delta_{min}/8+1, .., \delta_{max}/8\}$.

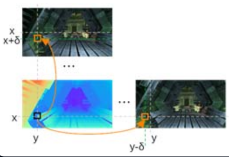

Fig 2: Horizontal and vertical disparity principle

Part 3 - Cost aggregation and disparity computation

Goal: Attributing a similarity score $S_r(i,j,\delta)$ to each pixel (i, j) and disparity candidate δ and compute disparity map.

Principle: 1) Apply six 3D Convolution blocks (see part 1), similar to [7] but with 64 output channels.
2) Final 3D convolution *layer*, with no normalization nor activation, 1 output channel, gives a final score $S_r(i,j,\delta)$.
3) Compute with soft argmax function on $S_r(i,j,\delta)$ a downsampled disparity map.
4) Upsample with a bilinear upsampling to upscale $\Delta_r{}^d$ to input resolution.

Output: Coarse disparity map

Part 4 - Multi-view disparity refinement

Goal: Refinement by reprojection of homologues color to obtain a refined disparity map Δ_r

Principle: 1) For each I_t, we compute the gathered image H, *i.e*, color values of the homologous pixels of I_r on I_t following the coarse disparity map.
2) Apply U-shaped residual 2D convolutional network [8] on H to obtain the residual disparity Δ . All of our layers have half as much output channels as [8] .
3) Compute refined disparity map with $\Delta_r = \Delta_r{}^u + \Delta_r{}^r$.

Output : Refined disparity map.

Training

Training set: We collected free-to-use 3D models and pictures which were randomly associated to create a 3D scene (see fig 4).
- 1 set = 16 FullHD images from 4x4 virtual camera array with their 16 disparity GT.
- 869 sets generated with δ in [0-270]
- 1 set gives 4 subsets (4 central views)
- The training loss is computed with both the coarse and fine output of the network

Iterations are not performed on the full images but on a random crops of 960 ×540.

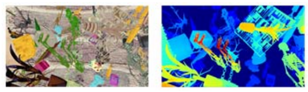

Fig 4 : Example from our training dataset, a reference image (left) and its disparity GT (right).

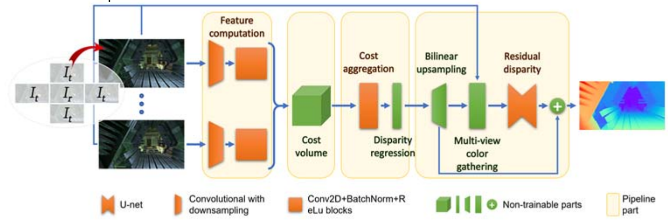

Fig 3: Overview of our 4-part solution: from the input (set of 4 RGB target images $\{I_t\}$ and reference image (I_r) to the generated disparity map of I_r. (Δ_r)

Results

Experiment conditions

Dataset: WLF hand designed test dataset of Li et al. [4].

Reference image : Only the central view.

The original camera array is 9×9, we only took as target images the top-middle, middle-right, bottom-middle and middle-left images of the array.

Metrics: bad 0.15 at bad 1 metrics for validation, i.e., percentage of disparity values above the bad threshold.

Hardware :
- CPU : Intel Xeon E5-2630 2.6Ghz
- GPU : NVIDIA Quadro RTX 5000 GPU 16Go

Timing on CPU : for [3], the code provided by the authors does not run on GPU and for [4], the network could not compute on GPU to the best of our efforts.

Method	Bad 0.15 (%)	Bad 0.3 (%)	Bad 0.6 (%)	Bad 1 (%)	Time (s)
[9]	37.79	**5.32**	**2.90**	**2.55**	600 *
[4]	15.04	7.05	3.95	2.80	40 *
[6]	25.71	10.67	4.15	3.22	1.6 **
Ours	**14.09**	8.13	4.89	3.30	**0.5****

Table 1: Results on the WLF [4] dataset compared to state of the art. Computation times were measured on (*)CPU, (**)GPU.

Method efficiency compared to state of art

- Inference time is much smaller, 3 times faster than [6].
- Higher precision measured by the quantification error (bad 0.15).

Limitations

- Thin and repetitive objects (e.g., vertical bars, like on a baby's bed).
- The method requires either adaptation or re-training for it to be efficient on edge and corner views of a camera array.

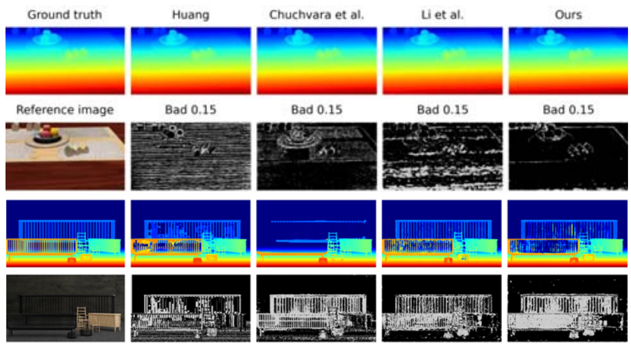

Fig 5: Disparity recovery for a reference image and quantification error (disparity error above 0.15). From left to right column: Disparity ground truth, RPRF approach [3], Chuchvara et al. [6], Li et al. [4] and our approach.

References

[1] Stéphanie Prévost, Cédric Niquin, Sylvie Chambon, and Guillaume Gales, Multi- and Stereoscopic Matching, Depth and Disparity, chapter 7, p. 137–155, John Wiley & Sons, Ltd, 2013.

[2] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," rXiv:2006.02535 [cs], Jun 2020, arXiv: 2006.02535.

[3] Chao-Tsung Huang, "Empirical bayesian light-field stereo matching by robust pseudo random field modeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 552–565, Mar 2019.

[4] Yan Li, Qiong Wang, Lu Zhang, and Gauthier Lafruit, "A lightweight depth estimation network for wide-baseline light fields," IEEE Transactions on Image Processing, vol. 30, pp. 2288–2300, 2021.

[5] Neus Sabater et al., "Dataset and pipeline for multi-view light-field video," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Jul 2017, p. 1743–1753, IEEE.

[6] Aleksandra Chuchvara, Attila Barsi, and Atanas Gotchev, "Fast and accurate depth estimation from sparse light fields," IEEE Transactions on Image Processing, vol. 29, pp. 2492–2506, 2020

[7] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," arXiv:1807.08865 [cs], Jul 2018, arXiv: 1807.08865

[8] Corinne Stucker and Konrad Schindler, "Resdepth: Learned residual stereo reconstruction," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Jun 2020, p. 707–716, IEEE.

[9] Jonathan T. Barron, "A general and adaptive robust loss function," in2019 IEEE/CVF Conference on ComputerVision and Pattern Recognition (CVPR). Jun 2019, p.4326–4334, IEEE

Université de Reims Champagne-Ardenne

EG'22 TheEvent