FACEBOOK AI

RidgeSfM: Structure from Motion via robust pairwise matching under depth uncertainty

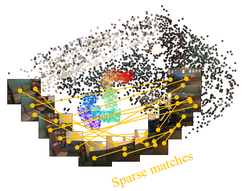
Benjamin Graham

David Novotny

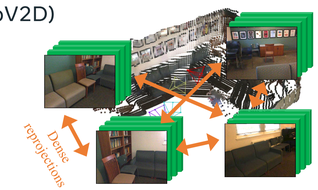
RidgeSfM

We consider the problem of simultaneously estimating a dense depth map and camera pose for a large set of images of an indoor scene.

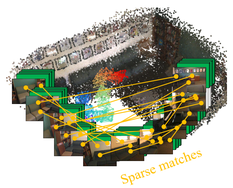
Classic SfM methods (COLMAP)
- Use sparse matching
- Generates sparse depth maps
- Scale to large numbers of views
- Does use global bundle adjustment
- Does not use deep learning



Modern deep SfM methods (BA-Net, DeepV2D)
- Use dense warping losses
- Use deep learning to ground the reconstruction with geometric prior
- High memory usage limits number of views that can be merged



RidgeSfM
- Use sparse matches
- Scales to large numbers of views
- Generates a dense reconstruction
- Uses deep learning
- Uses global bundle adjustment



Reconstruction pipeline overview

1. RidgeSfM applies a monocular depth prediction network to each image in the scene. The prediction for each image can be fine-tuned using a parameter $\beta \in \mathbb{R}^{32}$.
2. Sparse 2D keypoints are extracted from each image.
3. Pairwise-RidgeSfM matches keypoints across pairs of images:
   a) A small subset of the keypoints are picked (RANSAC-style).
   b) Keypoints are projected into 3D using the depth prediction.
   c) Camera orientation is estimated using Umeyama's algorithm
   d) Depth is fine-tuned conditional on the camera alignment.
   e) Iterate steps (c)-(d) optimizing alignment and depth predictions.
   f) Grow the set of matches as alignment improves.
4. Collect the sparse matches from Pairwise-RidgeSfM and perform a global bundle adjustment, fine-tuning the depth predictions and camera.

Depth prediction with fine-tuning

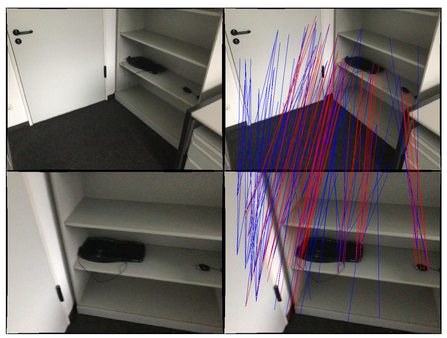
Monocular depth prediction ConvNet maps each RGB input image to
- a mean depth prediction, and
- factors of variation— a linear basis that spans the modes of uncertainty of the depth prediction.

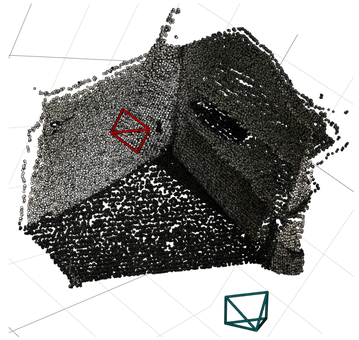
During training, use ridge regression to optimize a linear combination of the factors of variation to backpropagate through.

In use, adjust the dense depth planes based on sparse keypoint matches.

Top left: an input image.
Bottom left: the predicted depth.
Middle and right: We use SVD to reduce the 32 FoV planes down to 12 planes, and display them as 4 RGB images; each of the 4x3 color planes represents one factor of variation.

Pairwise-RidgeSfM: RANSAC style matching



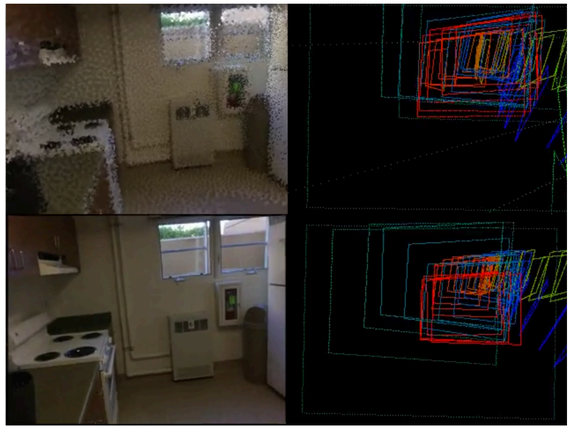


Example results on pairwise matching.
Top: Pairs of images with per-pixel correspondences.
Bottom: The inferred scene point clouds and cameras - we have plotted 10% of the pixels.
The blue lines show initial feature matches. Red matches denote the inliers of the Pairwise RidgeSfM alignment.

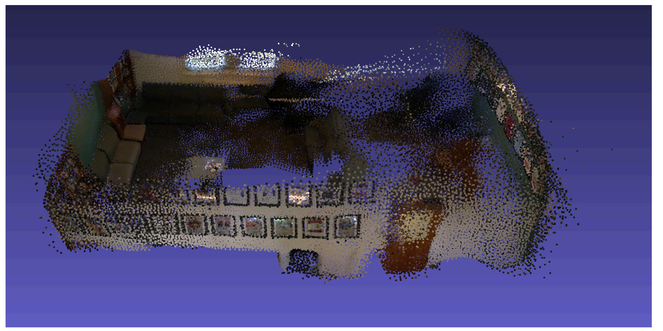
ScanNet scene reconstruction results

Method	COLMAP SfM pipeline				RidgeSfM using SuperPoint features			
Skip rate	1	3	10	30	1	3	10	30
Camera rotation (degrees)	22.12	11.17	9.85	29.85	7.09	7.84	7.35	12.84
Camera center (m)	0.973	0.597	0.540	1.085	0.296	0.314	0.331	0.489
Depth map L_1 err. (m)	0.941	0.763	0.727	1.184	0.221	0.234	0.243	0.322
Depth map RMSE (m)	1.138	1.012	0.990	1.386	0.305	0.332	0.343	0.432
PCL L_1 err. (m)	0.647	0.642	0.639	0.860	0.209	0.258	0.303	0.454
PCL RMSE (m)	0.821	0.885	0.906	1.081	0.289	0.345	0.393	0.569
Successful reconstructions	99%	100%	98%	81%	100%	100%	100%	100%

Quantitative comparison with COLMAP on large-scale bundle adjustment on the ScanNet dataset.
For COLMAP, evaluation is based on the available reconstructed frames for scenes where reconstruction was at least partially successful. For RidgeSfM, the evaluation uses all frames in and all scenes.



ScanNet scene 0707 reconstruction:
Top left: the rendered point cloud. Top right: The focal-plane trajectory for the predicted camera locations.
Bottom left: An input frame. Bottom right: The focal-plane trajectory of the ground truth camera locations.



Birds-eye view for the ScanNet scene 0708 reconstructed point cloud.

Take a photo to learn more:

https://github.com/facebookresearch/RidgeSfM