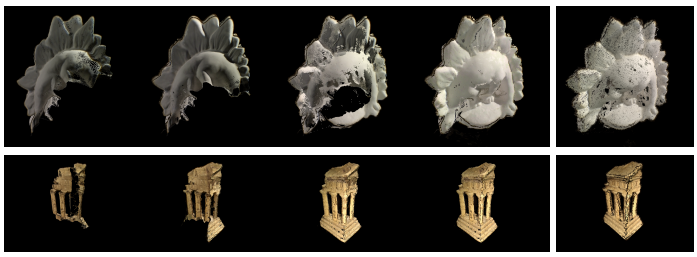
Next-Best Stereo: Extending Next-Best View Optimisation For Collaborative Sensors

Oscar Mendez[1]
O.Mendez@surrey.ac.uk
Simon Hadfield[1]
S.Hadfield@surrey.ac.uk
Nicolas Pugeault[2]
N.Pugeault@exeter.ac.uk
Richard Bowden[1]
R.Bowden@surrey.ac.uk

[1] Centre for Vision Speech and Signal Processing
University of Surrey Guildford, UK
[2] Department of Computer Science
University of Exeter
Exeter, UK



1 pair 3 pairs 5 pairs 7 pairs Reference (GT)

Figure 1: Results for Middlebury Dino (top) and Temple (bottom) Datasets, with varying numbers of stereo pairs. The final column shows the reference.

The 3D reconstruction of scenes and objects from 2D images is an extremely important part of many tasks, such as robot navigation, scene understanding and surveying. Mobile platforms with a single camera can easily overwhelm systems that attempt exhaustive optimisations over all images. Reconstruction algorithms that are capable of selecting data that maximises performance, while reducing computational time are necessary to perform reconstruction in the real world. This work proposes an approach to intelligently filter large amounts of data for 3D reconstructions of unknown scenes using monocular cameras. Figure 1 shows how the reconstruction progresses with a limited number of views. We can achieve state-of-the-art results using as little as 3.8% of the views on the Middlebury dataset. Furthermore, view selection is efficient, taking only 1.1ms per pose pair.

1 Next-Best View Optimisation

We first present a novel criterion for Next-Best View (NBV) optimisation based on a compromise between the competing objectives of coverage and accuracy. The coverage objective will drive the system to collect views of previously unobserved parts of the scene (e.g., due to restrictions on the field of view or occlusion), whereas the accuracy objective will drive the system to choose the next pose to reduce the point cloud's uncertainty. These two criteria are optimised jointly, making use of an octree structure and a dense point cloud. The octree allows for quick and efficient calculations on scene geometry, while the dense cloud (and covariances) allow for more detailed calculations about scene noise and and viewing angle.

The NBV is calculated as follows: Given a Configuration Space (CS) of sensor poses, the cost of each pose can be estimated by casting a set S_r of random rays from the camera centre through the image plane. Each ray will continue until it hits either an occupied (V_o) or unobserved (V_u) voxel, ignoring empty (V_e) voxels. When a ray $r \in S_r$ intersects with an occupied voxel $v \in V_o$, we can estimate a cost for each point $p \in P_v$ as $\phi(r,p) = e^{-||\lambda_p e_p \times r||}$, where λ_p and e_p are the largest eigenvalue and eigenvector, respectively, of the covariance Σ_p. Consequently, the cost of a voxel is defined as the average point cost

$$\psi(r,v) = \frac{1}{|P_v|} \sum_{p \in P_v} \phi(r,p). \tag{1}$$

The NBV cost of a particular pose x is defined as

$$C_x = \frac{1}{|S_r|} \sum_{r \in S_r} \begin{cases} \psi(r,v) & \text{if } v \in V_o \\ \gamma \in [0,1] & \text{else } v \in V_u. \end{cases} \tag{2}$$

In this equation, γ is a parameter that can encourage or discourage exploration. As shown in Figure 2 γ of 1 will give the highest cost to unobserved voxels, prefering to reduce the uncertainty of observed voxels, while 0 will give them the lowest, preferring exploratory behaviour.

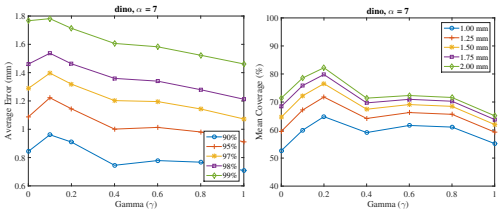


Figure 2: Average Error (Left) and Average Coverage (Right) with different values of γ.

2 Next-Best Stereo Optimisation

When there are multiple collaborating sensors available, we can extend NBV to also optimise the stereo arrangement of the sensors. This can be achieved by selecting another view, with respect to the NBV, to create the best possible stereo pair. Actively selecting stereo pairs allows sensors to be positioned to allow an optimal vergence and baseline, respective to the observed parts of the scene.

This implies several requirements: First, the baseline of the cameras must be scaled, depending on the distance to the observed geometry, and the vergence angle should be minimised to allow the dense matching to be performed with the least amount of error possible (C_B, C_T). Second, we must ensure robustness against rotation in the image (C_R). Finally, the distance between the vergence point and the nearest geometry should be minimised, to ensure that the sensors are trained on actual scene geometry and not empty space (C_G). We then find the pose that minimises

$$C = C_B + C_T + C_R + C_G. \tag{3}$$

Please see full paper for more details.

3 Results

Table 1 shows experimental evaluation against the Middlebury benchmark. It can be seen that the proposed method allows efficient selection of stereo pairs for reconstruction, such that a dense model can be obtained with only a small number of images. Once a complete model has been obtained, the remaining computational budget is used to intelligently refine areas of uncertainty, achieving results comparable to state-of-the-art batch approaches on the Middlebury dataset, using as little as 3.8% of the views.

	Thresh.	[1]	[1]	[2]	Proposed
Num. Frames	-	41	41	unknown	**26**
Error (mm)	80%	0.64	0.59	0.64	**0.53**
	90%	1.00	0.88	0.91	**0.74**
	99%	2.86	2.08	1.89	**1.68**
Coverage (%)	0.75mm	79.5	82.9	72.9	**87.3**
	1.25mm	90.2	93.0	73.8	**96.4**
	1.75 mm	94.3	96.9	73.9	**98.4**

Table 1: Middlebury Evaluation for different NBV and MVS approaches.

Both contributions are extremely efficient, taking 0.8ms and 0.3ms per pose, respectively. More importantly, neither uses any image-based information instead relying on cues from the partially reconstructed geometry. This allows the proposed approach to sample areas of space that have not been imaged, and is therefore inherently applicable to robotic problems such as path-planning and goal estimation.

[1] Alexander Hornung, Boyi Zeng, and Leif Kobbelt. Image selection for improved multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.

[2] Michal Jancosek, Alexander Shekhovtsov, and Tomas Pajdla. Scalable multi-view stereo. In *International Conference on Computer Vision*, 2009. ISBN 9781424444410.