CVSSP Centre for Vision, Speech and Signal Processing

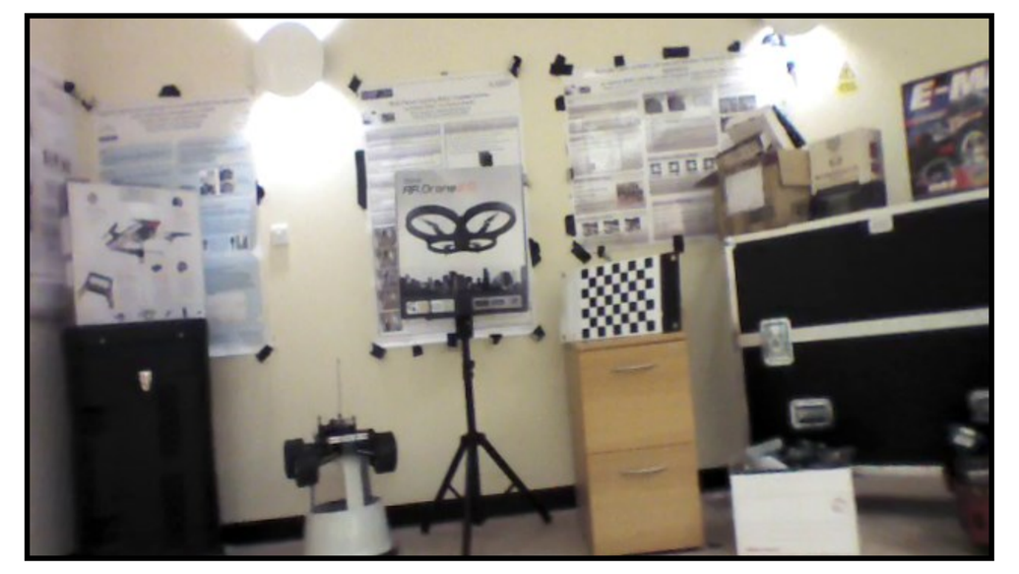
Exploiting high level scene cues in stereo reconstruction

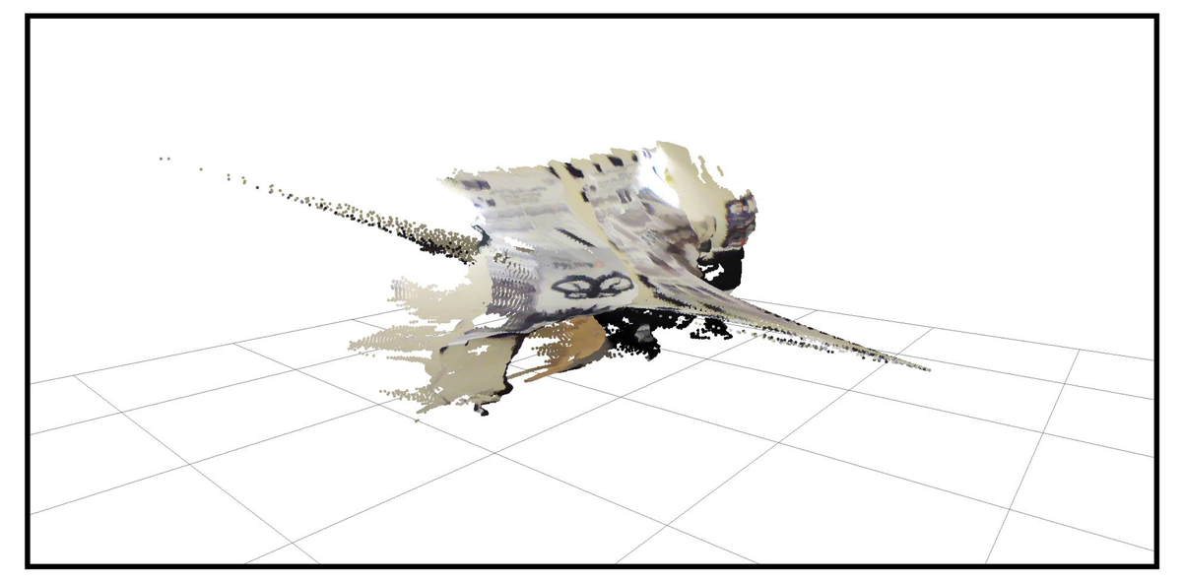
Simon Hadfield (S.Hadfield@surrey.ac.uk) Richard Bowden (R.Bowden@surrey.ac.uk)

UNIVERSITY OF SURREY

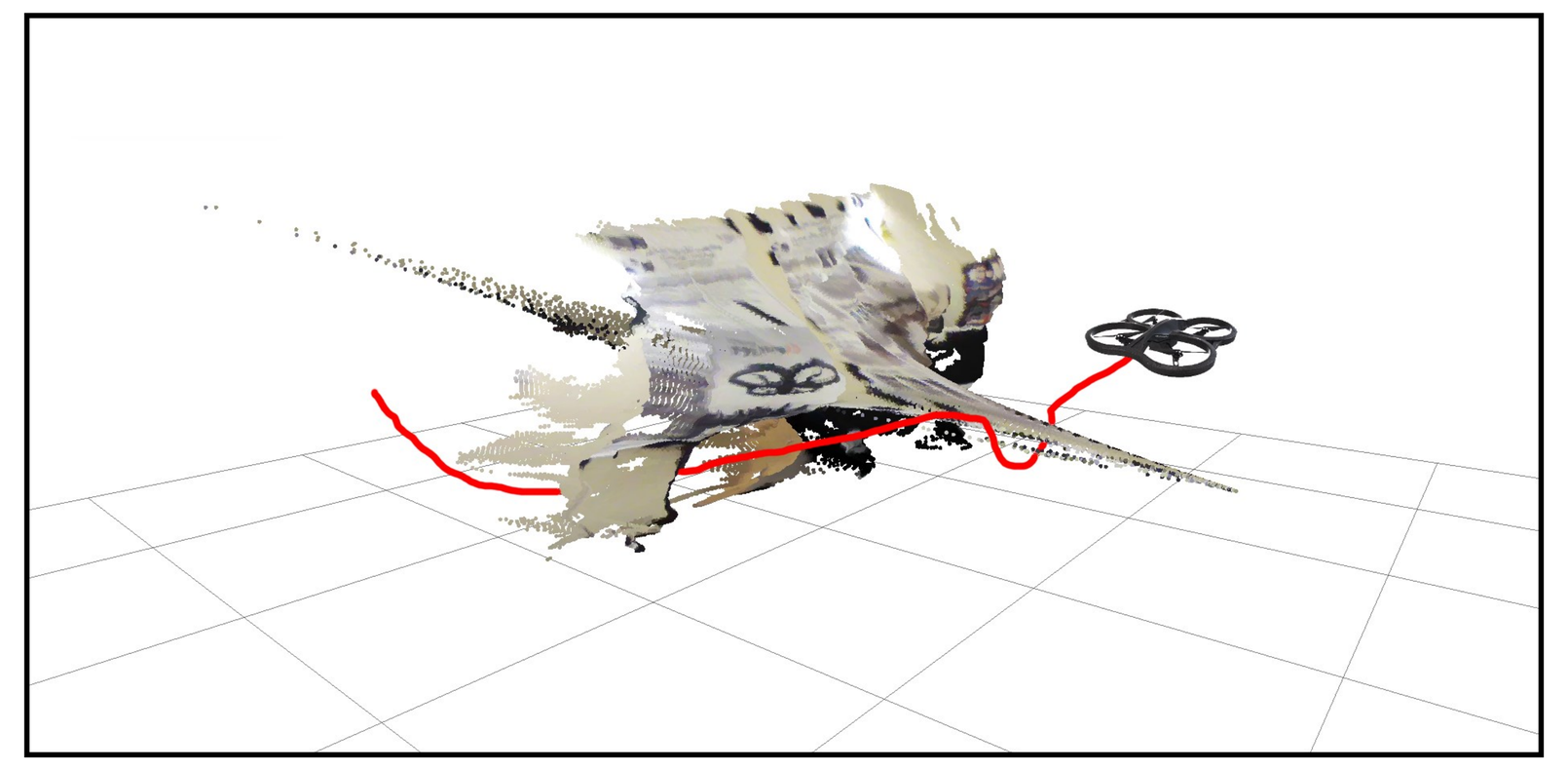
Robust stereo reconstruction

- When reconstruction goes wrong, it goes **really** wrong





- Matching failures generate totally unrealistic scenes

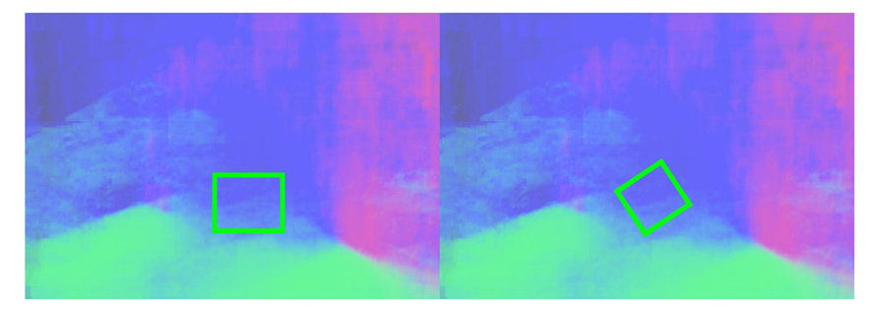


- Humans don't have this problem
- They can even get 3D without matching, from a single photo
- **They understand the scene**

Top-down understanding cues

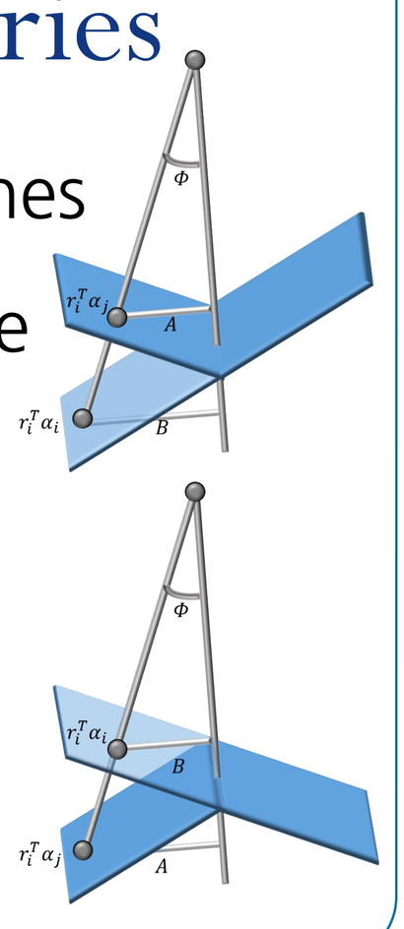
Normal matching

- As bottom up costs, using estimated surface normals
- Rotate to same co-ordinate frame $I_n^r(G(x^r)) - RI_n^t(G(H(x^r \mid \alpha)))$



Edge categories

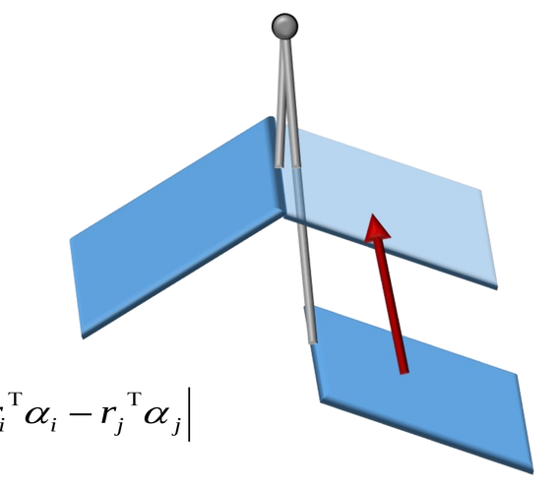
- Boundary between planes
- If recognised as concave
  - ◊ Favour B > A
- If recognised as convex
  - ◊ Favour A > B



$$\sum_{N_i} \left|\sin(\phi_i)(r^T\alpha_i - r^T\alpha_j)\right| + \sum_{N_j} \left|\sin(\phi_j)(r^T\alpha_i - r^T\alpha_j)\right|$$

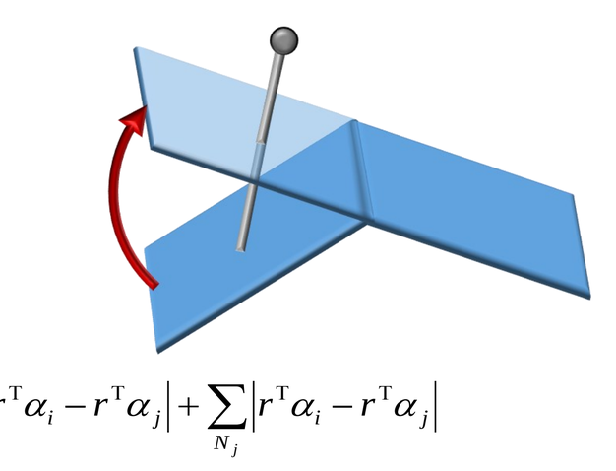
Connected

- Applies when not an occlusion
- Favours 3D connected edge $\sum_{N_{i,j}} \left|r_i^T\alpha_i - r_j^T\alpha_j\right|$



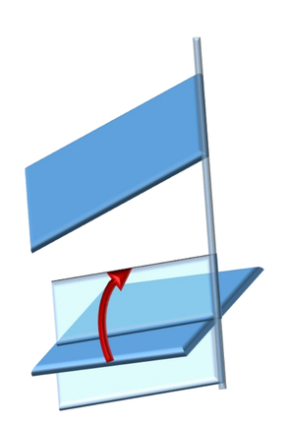
Coplanar

- Applies when not recognised as an edge
- Favours a single larger plane $\sum_{N_i} \left|r^T\alpha_i - r^T\alpha_j\right| + \sum_{N_j} \left|r^T\alpha_i - r^T\alpha_j\right|$



Collinear

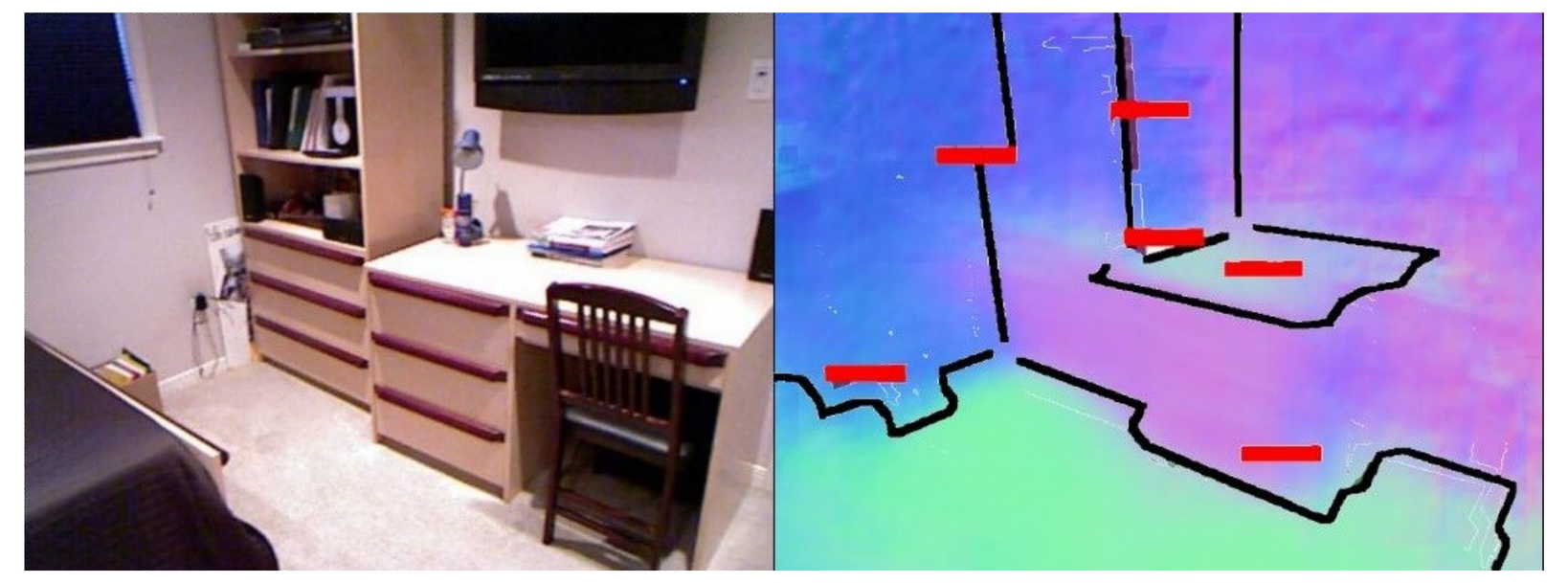
- Two non-neighbouring planes
- Applies when edges collinear in image
- Favour edges collinear in 3D



$$\sum_{N_{i,e}} \left|r^T\alpha_i - r^T\alpha_j\right| + \sum_{N_{j,e}} \left|r^T\alpha_i - r^T\alpha_j\right|$$

Scene understanding

- Estimate scene geometry
- Edges separating dominant planes [1]
  - ◊ Concave, Convex & Occlusion
- Surface normals [2]

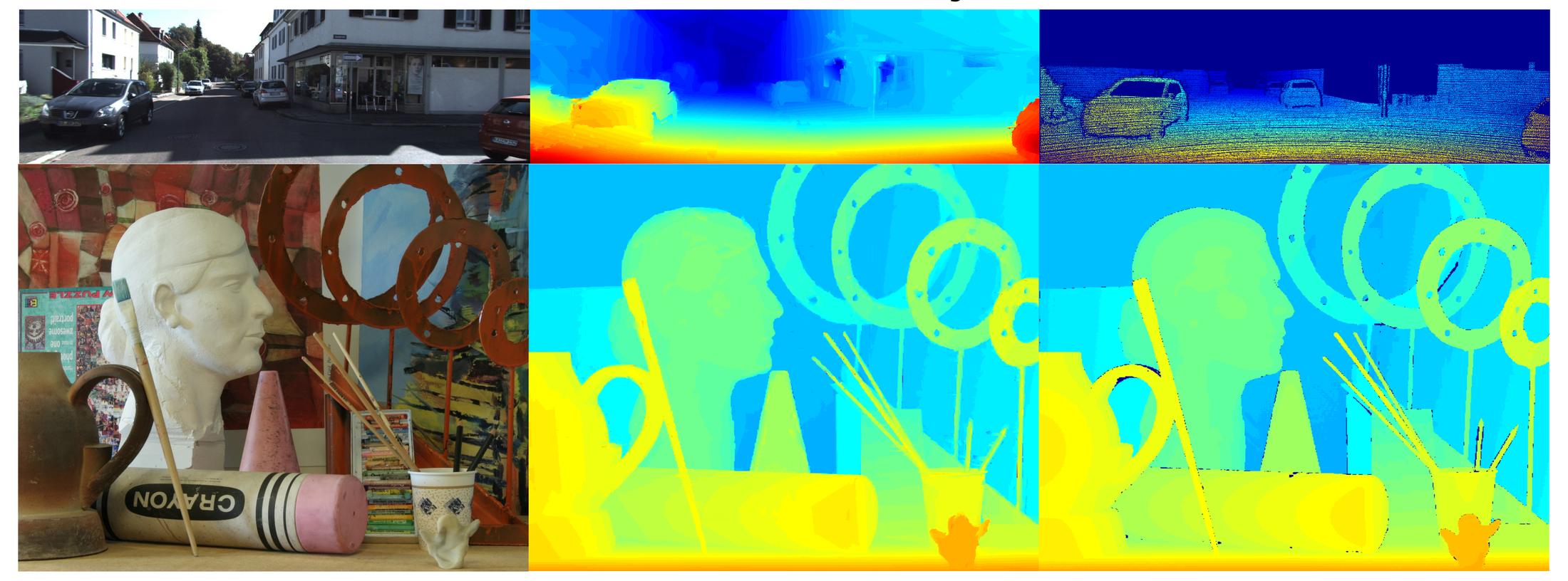


Unified framework

- Perform superpixel segmentation on reference image
- Reconstruct a slanted 3D plane per superpixel
- Penalise fractional depth error
  - ◊ Helps balance types of cue based on scene
- All cues are linear in plane parameters (α)
  - ◊ However, some are iterative approximations
- Iterative extension to L1 optimiser [3]

Results

- Results on KITTI and Middlebury 2014

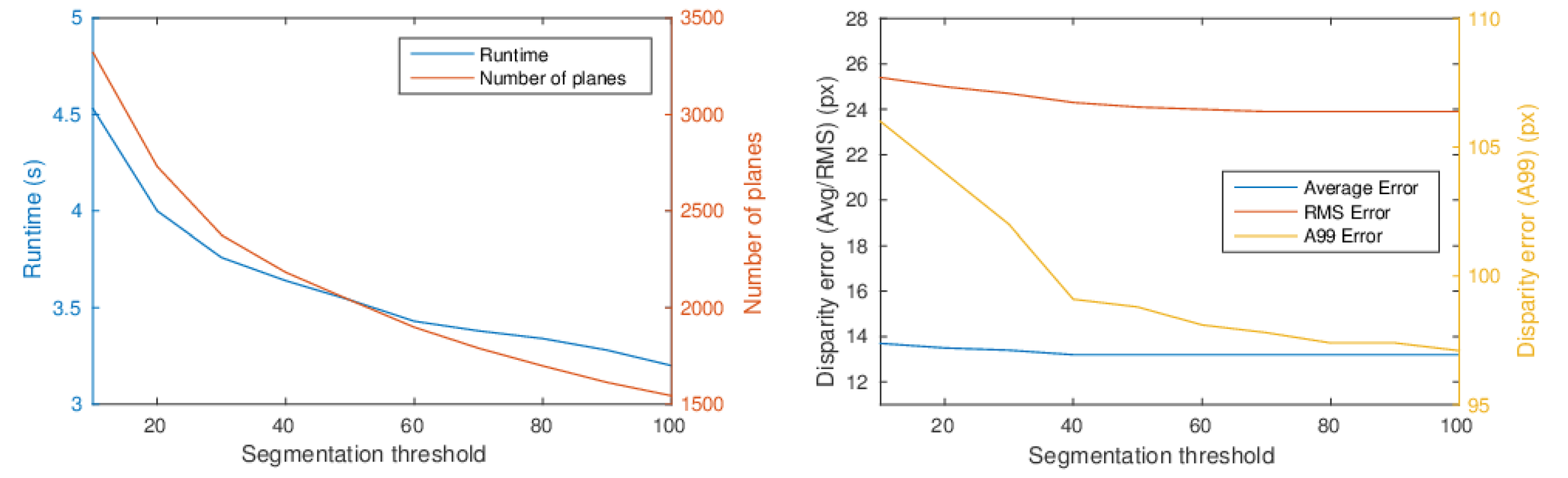


- Measured accuracy: RMS & Average disparity error
- Measured robustness (lack of outliers): A99 error

Technique	RMS Err. (px)	Avg. Err. (px)	A99 (px)	Time (s)
MC-CNN	4.49	1.93	20.1	**3.59**
MC-CNN + HLSC	3.77	1.41	17.3	310
Mesh Stereo	6.78	2.48	33.6	376
Mesh Stereo + HLSC	6.56	1.99	32.5	526
CoR	3.82	1.33	17.6	17.32
CoR + HLSC	**3.43**	**0.98**	**16.8**	150

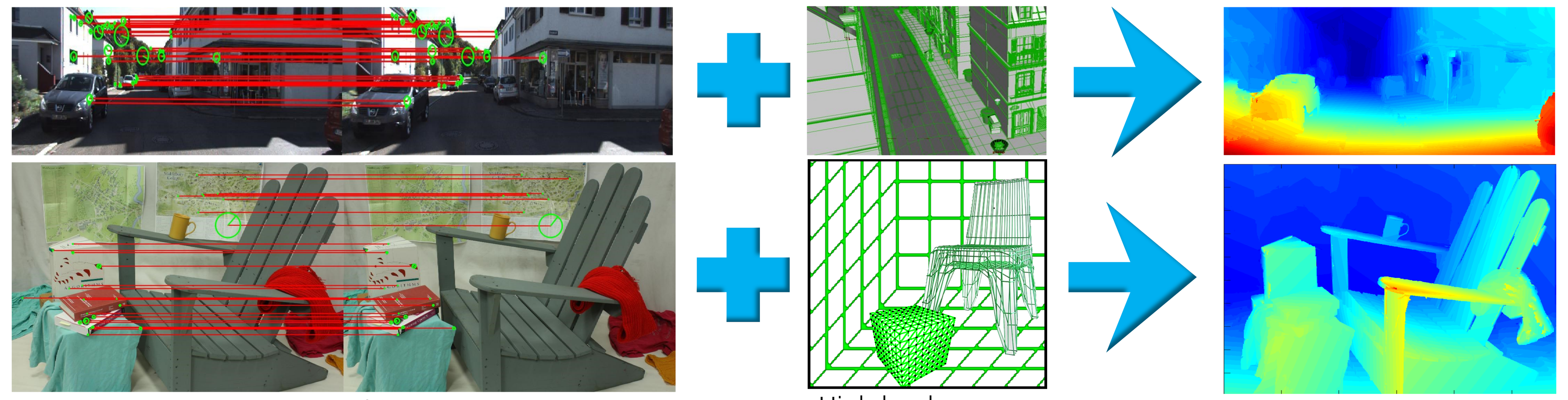
- Up to **30% improvement** using High level cues

Omitted cue	RMS Err. (px)	Avg. Err. (px)	A99 (px)	Time (s)
Surface Normal	25.1	13.1	101	140
Connected	24.7	12.7	121	151
Coplanar	24.9	12.7	117	139
Collinear	23.8	10.3	98	148
Edge category	24.0	10.1	103	130
Brightness Const.	35.7	19.3	138	130
Gradient Const.	31.3	12.6	121	125
Census	28.8	10.9	109	**112**
Triangulation	43.4	22.1	145	163
-	**23.3**	**9.9**	**94**	161



Contributions

- Unify *Scene Understanding*, with *Stereo Reconstruction*
- Integrate many cues from both fields
- Automatically balance cue types based on the scene



Stereo matching High level cues

- **Bottom-up + Top-down = Improved Reconstruction**

Bottom-up matching cues

Brightness Constancy

- Planes induce homographies based on orientation & position

- Assume appearance same from any view (brightness constancy)

$$I^r(G(x^r)) - I^t(G(H(x^r \mid \alpha)))$$

- Linearise with Taylor expansion
- Obtain *Optical Flow Constraint* in terms of plane parameters

$$I^r(G(x^r)) - I^t(G(H(x^r \mid \alpha^0))) - J_I J_G J_H \Delta\alpha$$

Gradient Constancy

- Apply the same cost on gradient images

$$I_\Delta^r(G(x^r)) - I_\Delta^t(G(H(x^r \mid \alpha^0)))$$

Census Matching

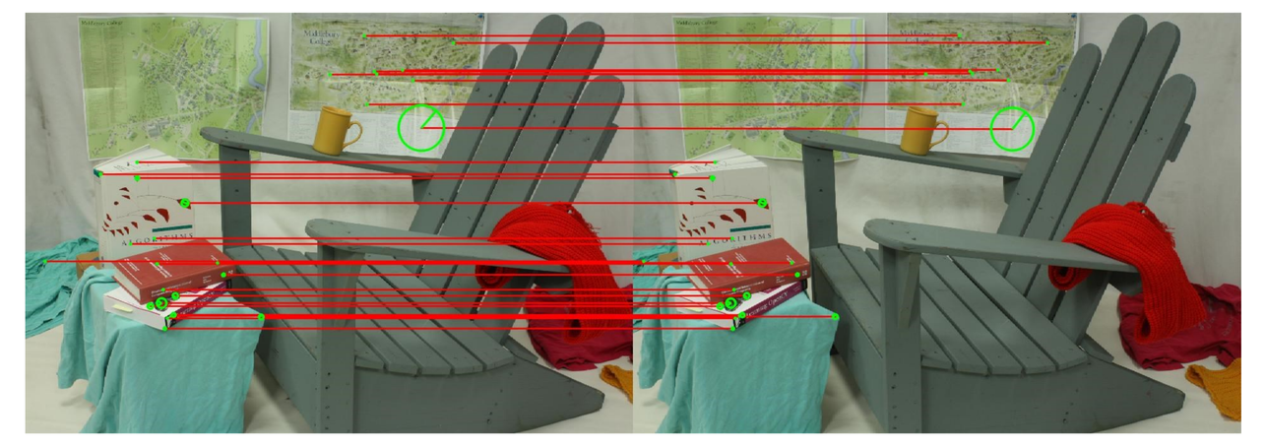
- For Census images, use Hamming distance (XOR)

$$I_C^r(G(x^r)) \oplus I_C^t(G(H(x^r \mid \alpha)))$$

Triangulation

- Use matches to triangulate 3D points
- Favour planes close to 3D points
- Default uses CNN matches, can use existing bottom-up stereo

$$r^T\alpha\hat{d} - 1$$



References

[1] D. Fouhey, A. Gupta, M. Hebert, **Unfolding an indoor Origami world**, in ECCV 2014.

[2] D. Fouhey, A. Gupta, M. Hebert, **Data-driven 3D primitives for single image understanding**, in ICCV 2013.

[3] A. Saxena, M. Sun, A. Ng, **Make3D: Learning 3-D Scene Structure from a Single Still Image**, In PAMI 2008.

Acknowledgements

This work was funding by EPSRC grant "Learning to recognise dynamic visual content from broadcast footage" (EP/I011811/1).