Supplementary material - Exploiting high level scene cues in stereo reconstruction

Simon Hadfield
University of Surrey
Guildford, UK, GU2 7XH
s.hadfield@surrey.ac.uk

Richard Bowden
University of Surrey
Guildford, UK, GU2 7XH
r.bowden@surrey.ac.uk

A. Linearisation of appearance matching costs

As mentioned in Section 6 of the paper, all the cost functions integrated into the system are linear in terms of $\boldsymbol{\alpha}$, except for the image lookups during appearance matching (Equations 3-5 of the paper). Here, we describe how these lookups are linearised, using an extension of the "Optical flow constraint" from the motion estimation literature. This enables us to perform efficient inference by solving a sparse Linear Program.

We illustrate the linearisation for the Brightness Constancy cost function

$$E_{bc}\left(s_i\right) = \sum_{\mathbf{x}_i^r \in s_i} \psi\left(\mathrm{I}^r(\mathbf{x}_i^r) - \mathrm{I}^t(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right))\right), \quad (1)$$

however, it extends trivially to the other appearance matching costs.

As mentioned at the end of Section 4, this equation omits the conversion from 3 element homogeneous pixel positions ($\mathbf{x}$) to 2D image locations ($\tilde{\mathbf{x}}$). We now include a conversion

$$\tilde{\mathbf{x}} = \mathrm{G}(\mathbf{x}) \quad (2)$$

into the cost function explicitly

$$E_{bc}\left(s_i\right) = \sum_{\mathbf{x}_i^r \in s_i} \psi\left(\mathrm{I}^r(\mathrm{G}\left(\mathbf{x}_i^r\right)) - \mathrm{I}^t(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right)\right))\right). \quad (3)$$

The lookup in the reference image (I^r) does not depend on $\boldsymbol{\alpha}$ and so does not need to be linearised. For the target image (I^t) lookup, we perform a Taylor expansion and drop all terms of quadratic order or higher. If we have a current estimate $\boldsymbol{\alpha}^0$, we can perform the Taylor expansion around this and obtain a parameter update $\Delta\boldsymbol{\alpha}$

$$\mathrm{I}^t\big(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i^0 + \Delta\boldsymbol{\alpha}\right)\right)\big) \approx \mathrm{I}^t\big(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i^0\right)\right)\big) + \mathbf{J}\Delta\boldsymbol{\alpha}, \quad (4)$$

where $\mathbf{J}$ is the Jacobian of the combined function.

The function being approximated can be viewed as the composition of 3 functions (I^t, G and H). Thus, we can compute $\mathbf{J}$ in closed form using the total derivative chain rule,

$$\mathbf{J} = \mathbf{J}_I(\mathrm{G}\left(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right)\right))\, \mathbf{J}_G(\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right))\, \mathbf{J}_H(\mathbf{x}_i^r). \quad (5)$$

In other words, $\mathbf{J}$ is the matrix product of the three Jacobians ($\mathbf{J}_I$, $\mathbf{J}_G$ and $\mathbf{J}_H$) for the composited sub-functions, with each Jacobian being evaluated at the location output by its preceding sub-functions.

To define the first sub-Jacobian ($\mathbf{J}_H \in \mathbb{R}^{3\times3}$), remember that the H function is defined as the application of 3 matrix multiplications to the pixel position. First the inverse of the calibration matrix for the reference camera, second the homography matrix induced by the plane, and finally the intrinsic matrix for the target camera

$$\mathrm{H}\left(\mathbf{x}_i^r|\boldsymbol{\alpha}_i\right) = \mathbf{K}_t\mathbf{H}_i\mathbf{K}_r^{-1}\mathbf{x}^r = \mathbf{x}^t. \quad (6)$$

The matrix $\mathbf{H}_i$ is defined as $\mathbf{H}_i = \mathbf{R} + \mathbf{t}\boldsymbol{\alpha}_i^\top$ and is the only part of the equation which depends on $\boldsymbol{\alpha}$. As such, the sub-jacobian $\mathbf{J}_H$ in terms of $\boldsymbol{\alpha}$ is given by

$$\mathbf{J}_H(\mathbf{x}^r) = \mathbf{K}_t\mathbf{t}(\mathbf{K}_r^{-1}\mathbf{x}^r)^\top \quad (7)$$

the intrinsics of the target camera, and the outer product of the camera baseline with the normalised homogeneous pixel position.

The second sub-Jacobian ($\mathbf{J}_G \in \mathbb{R}^{2\times3}$) is the simplest, given by

$$\mathbf{J}_G(\mathbf{x}^t) = \begin{bmatrix} \frac{1}{w} & 0 & -\frac{u}{w^2} \\ 0 & \frac{1}{w} & -\frac{v}{w^2} \end{bmatrix} \quad (8)$$

where u, v, w are the elements of $\mathbf{x}^t$.

The final sub-Jacobian $\mathbf{J}_I \in \mathbb{R}^{1\times2}$ encodes how the image intensity varies as a result of changes in the pixel position, and is constructed from the x and y gradients of the target image.

$$\mathbf{J}_I(\tilde{\mathbf{x}}_i^t) = \mathrm{I}_\Delta^t(\tilde{\mathbf{x}}_i^t) \quad (9)$$

	Res.	Adirondack	ArtL	Jadeplant	Motorcycle	MotorcycleE	Piano	PianoL	Pipes
RMS Err.	F	13.1 / 1	17.6 / 2	207 / 6	21.3 / 5	21.0 / 3	11.1 / 2	14.3 / 2	27.9 / 2
	H	11.9 / 3	20.8 / 6	81.0 / 3	22.6 / 7	22.5 / 5	11.3 / 3	19.3 / 3	31.5 / 4
	Q	13.3 / 2	21.3 / 2	75.4 / 2	26.2 / 2	25.9 / 2	12.2 / 1	15.3 / 1	36.7 / 2
Avg. Err.	F	5.91 / 3	8.04 / 4	146 / 6	9.36 / 5	9.00 / 3	7.90 / 5	9.36 / 2	13.0 / 4
	H	5.76 / 7	9.91 / 6	39.6 / 6	9.85 / 7	9.54 / 5	8.09 / 7	10.8 / 4	15.0 / 8
	Q	6.75 / 2	11.7 / 2	38.6 / 2	11.5 / 4	11.2 / 2	8.58 / 2	10.3 / 2	18.7 / 4
A99	F	66.5 / 1	78.0 / 1	533 / 6	123 / 3	122 / 3	39.8 / 1	47.7 / 1	123 / 2
	H	58.0 / 2	84.6 / 6	319 / 2	132 / 5	132 / 3	38.1 / 1	100 / 4	131 / 4
	Q	63.3 / 2	75.2 / 1	283 / 1	140 / 2	139 / 2	43.6 / 1	56.2 / 1	149 / 2

	Res.	Playroom	Playtable	PlaytableP	Recycle	Shelves	Teddy	Vintage	**Average**
RMS Err.	F	30.6 / 5	36.7 / 3	18.3 / 4	10.0 / 2	21.7 / 3	15.2 / 5	155 / 6	39.3 / 5
	H	26.8 / 7	42.6 / 3	18.7 / 7	9.63 / 2	21.5 / 7	10.8 / 5	28.3 / 4	24.8 / 3
	Q	18.6 / 2	22.4 / 1	19.5 / 2	11.6 / 2	20.0 / 2	8.66 / 2	22.7 / 1	24.0 / 2
Avg. Err.	F	12.4 / 5	24.8 / 5	13.6 / 6	6.87 / 6	13.3 / 5	4.50 / 5	91.3 / 6	24.0 / 6
	H	11.0 / 7	25.2 / 4	14.0 / 9	6.74 / 9	13.3 / 7	2.97 / 5	19.5 / 9	12.9 / 6
	Q	9.47 / 2	15.8 / 3	14.6 / 4	7.64 / 4	12.4 / 2	3.73 / 2	16.8 / 3	13.2 / 2
A99	F	164 / 5	140 / 1	54.8 / 4	34.7 / 1	85.5 / 2	74.0 / 5	427 / 6	134 / 5
	H	147 / 7	184 / 3	56.9 / 4	37.5 / 1	79.4 / 5	40.6 / 2	81.2 / 1	106 / 2
	Q	99.9 / 2	77.9 / 1	59.5 / 2	43.0 / 1	79.2 / 1	43.5 / 2	61.7 / 1	98.1 / 1

Table 1: Performance on each sequence, for all resolution benchmarks. Listed is the error value for that sequence, followed by the ranking for that sequence. Rankings are out of 6, 9 and 5 for the F, H and Q benchmarks respectively.

Given these definitions, we can substitute the approximation of Equation 4 into the cost function from Equation 3

$$E_{bc}(s_i) = \sum_{\mathbf{x}_i^r \in s_i} \psi(\mathrm{I}^r(\mathrm{G}(\mathbf{x}_i^r)) - \mathrm{I}^t(\mathrm{G}(\mathrm{H}(\mathbf{x}_i^r | \boldsymbol{\alpha}_i^0))) - \mathbf{J}_I \mathbf{J}_G \mathbf{J}_H \Delta\boldsymbol{\alpha}). \quad (10)$$

This cost function is linear in terms of $\Delta\boldsymbol{\alpha}$.

B. Additional results

For the Middlebury 2014 dataset, the full breakdown of the performance across every sequence is given in table 1. This includes results for all 3 resolution benchmarks. Qualitative examples of the algorithm's output are given in figures figures 1 to 3. These examples were randomly chosen from the Middlebury 2014 and KITTI datasets.

C. Surface Normal Estimation

We performed an additional analysis of the surface normal estimation system. The agreement of the surface normals estimated in the 2 views was quantified by

$$\frac{\mathbf{R}\boldsymbol{I}_s^r(\mathbf{x}_i^r) \cdot \boldsymbol{I}_s^t(\tilde{\mathbf{x}}_i^t) + 1}{2}, \quad (11)$$

where $\tilde{\mathbf{x}}_i^t$ is the corresponding pixel in the target image according to the ground truth disparity map. In other words, the dot product of the 2 normal vectors in the target coordinate frame, normalised between 0 and 1.

Figure 1: Random examples from the Middlebury 2014 Full resolution benchmark. Input image (left) and output (right).

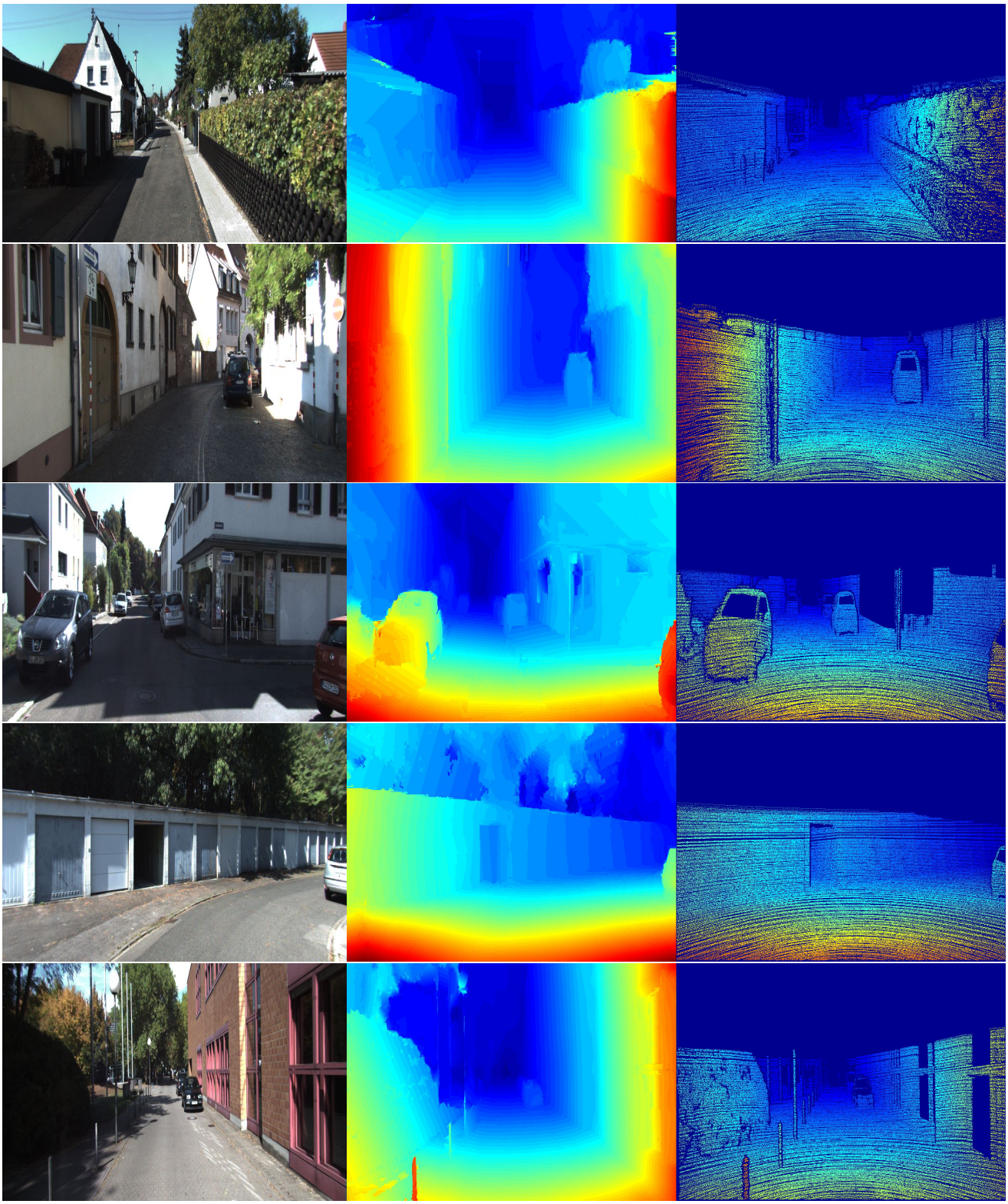

Figure 2: Random examples from the KITTI dataset. Input image (left), the output of our algorithm (middle) and the ground truth (right).

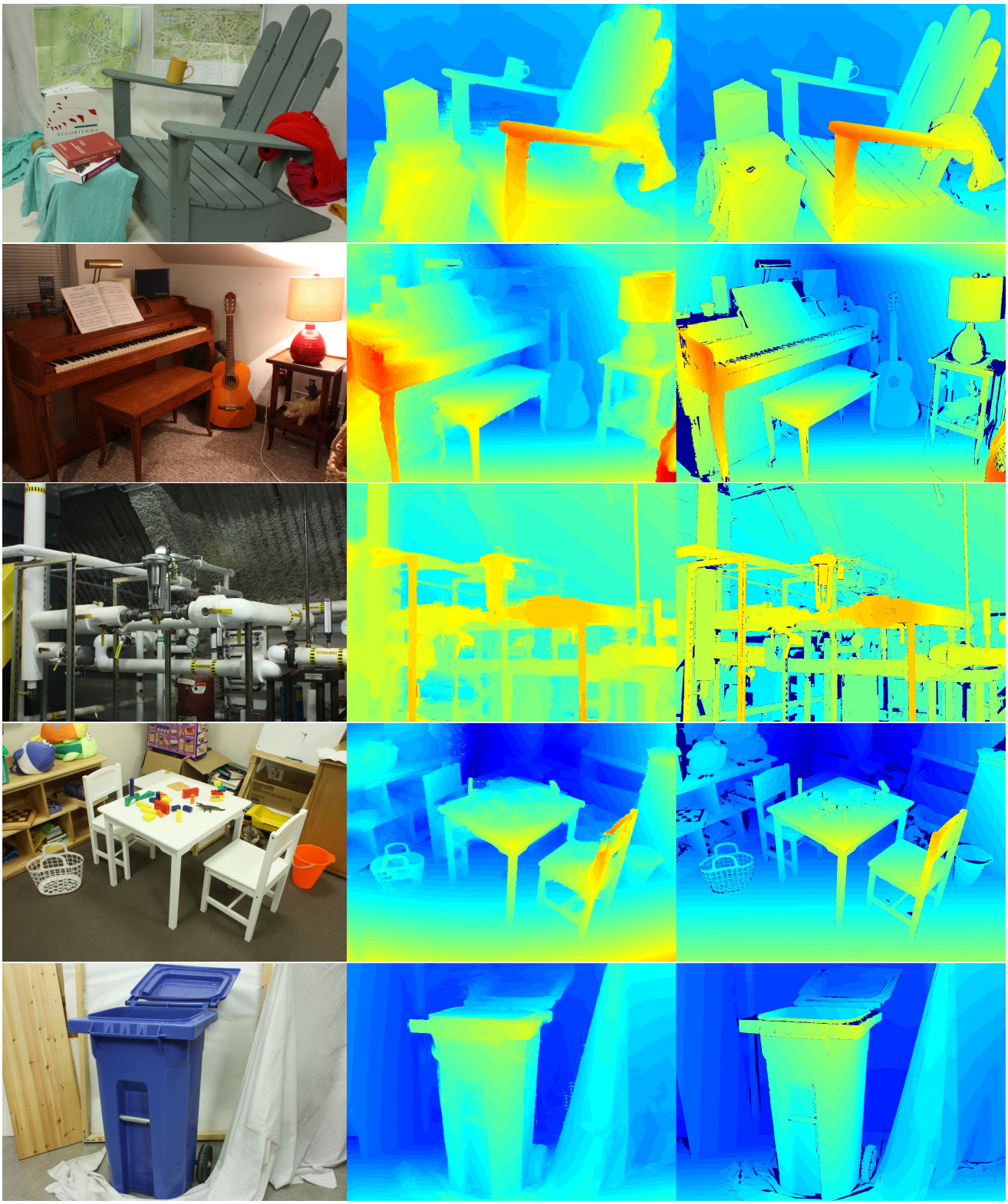

Figure 3: Random examples from the Middlebury 2014 Full resolution benchmark. Input image (left), the output of our algorithm (middle) and the ground truth (right).

The distribution across all pixels from all scenes in the Middlebury 2014 training set, is displayed in figure 4. The average consistency across viewpoints is 0.982, with almost nothing below 0.95. To measure the sensitivity of the overall reconstruction framework to this, we also computed the correlation between the surface normal agreement and the disparity error (again over all pixels in the training sequences). The resulting correlation co-efficient was -0.08, indicating a slight anti-correlation (i.e. increased surface normal agreement indicates a reduction in disparity error). This is reasonable as matching estimated surface normals is one of the inputs to the system, however the sensitivity proves very slight due to the influence of other cues.

D. Performance vs. baseline

This experiment examines the effect of the stereo baseline on the performance of the proposed system. The Middlebury 2014 and KITTI datasets are poorly suited for this evaluation as there is little variation in baseline, and the change in scene clutter is far more significant. Instead we use the Middlebury 2003 dataset which includes a larger array of cameras. We can then use different pairs of images to simulate stereo pairs with different baselines, but all viewing the same scene. The results are plotted in figure 5, which shows that an extremely narrow baseline is the most detrimental, and that good performance can be obtained for a wide range of baselines. However, there is an eventual decay in performance when the baseline becomes too large.

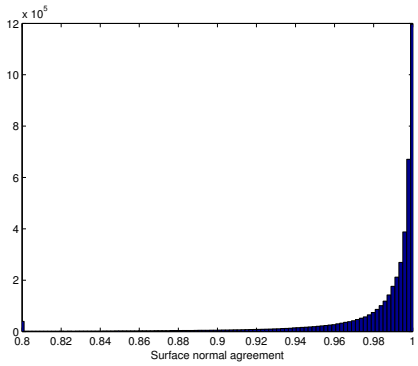

Figure 4: Distribution of consistency between the two viewpoints for estimated surface normals.

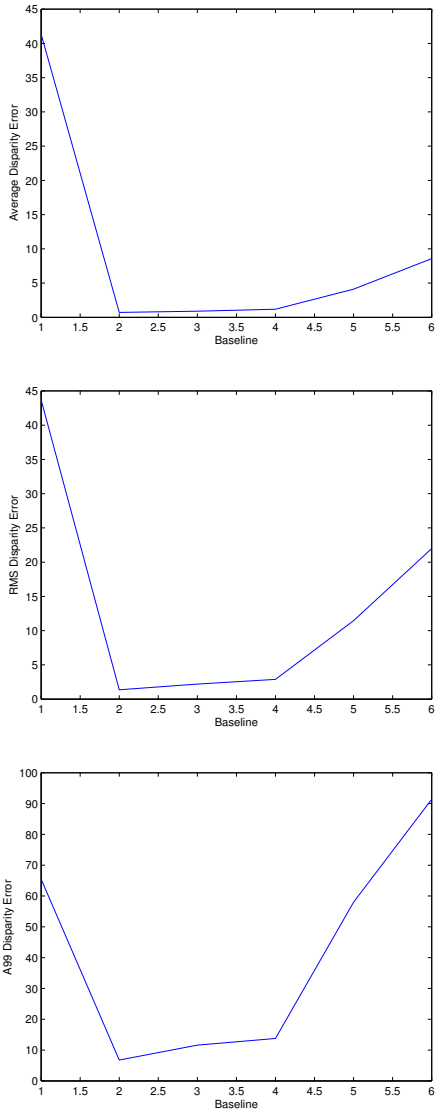

Figure 5: Performance against varying stereo baseline.