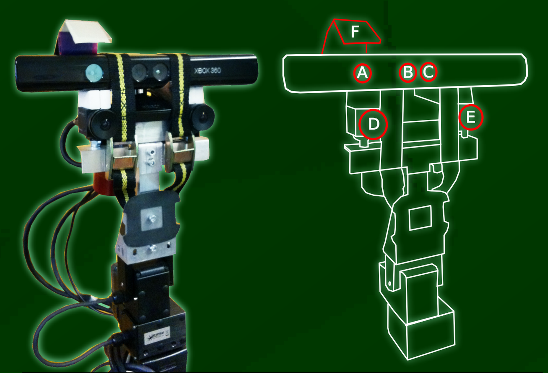
Teaching Stereo Perception to YOUR Robot

Marcus Wallenberg, Per-Erik Forssén
{wallenberg, perfo}@isy.liu.se
Computer Vision Laboratory, Linköping University

Overview

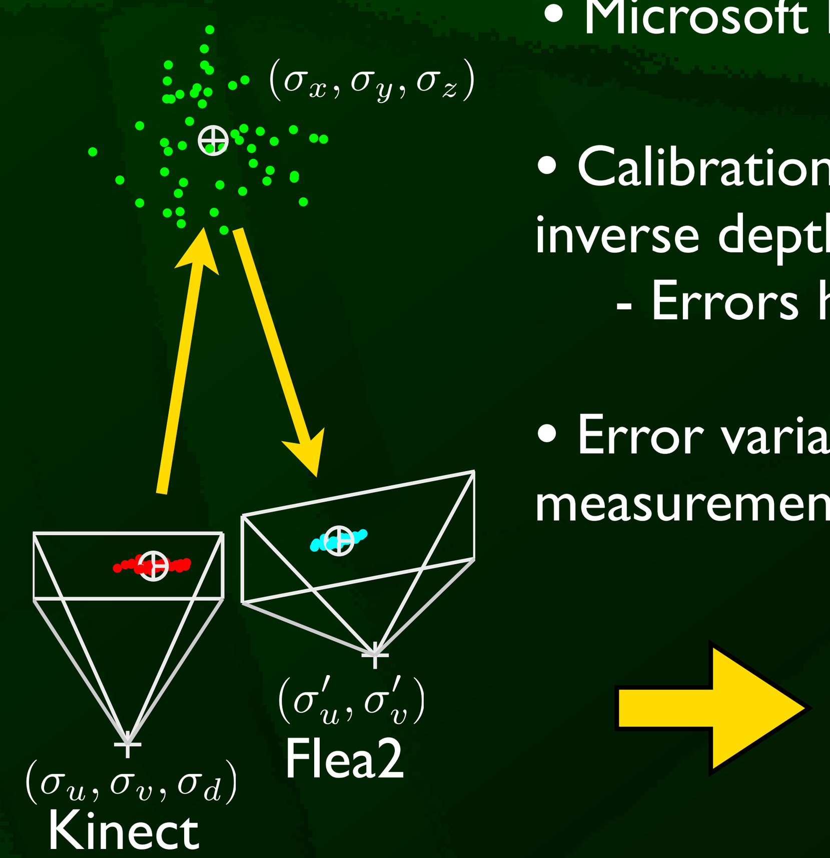
Object recognition on robot platforms is greatly facilitated by stereo vision. Wide-angle stereo is especially useful, since it provides a scene overview even at short range.
Traditional stereo vision using using rectification and alignment of images is often impractical on wide-angle images.
For best results, stereo should be adapted to take both hardware and scene characteristics into account. This implies that ground-truth acquisition for the target platform is desireable.
Here [1], a ground-truth acquisition and tuning procedure is used to automatically tune an extension to the *best-first propagation* (BFP) [2] algorithm.
The tuned correspondence algorithm is evaluated in terms of accuracy, robustness and ability to generalise. Both the tuning cost function and the evaluation are designed to balance the accuracy-robustness trade-off inherent in patch-based methods.

Hardware and calibration



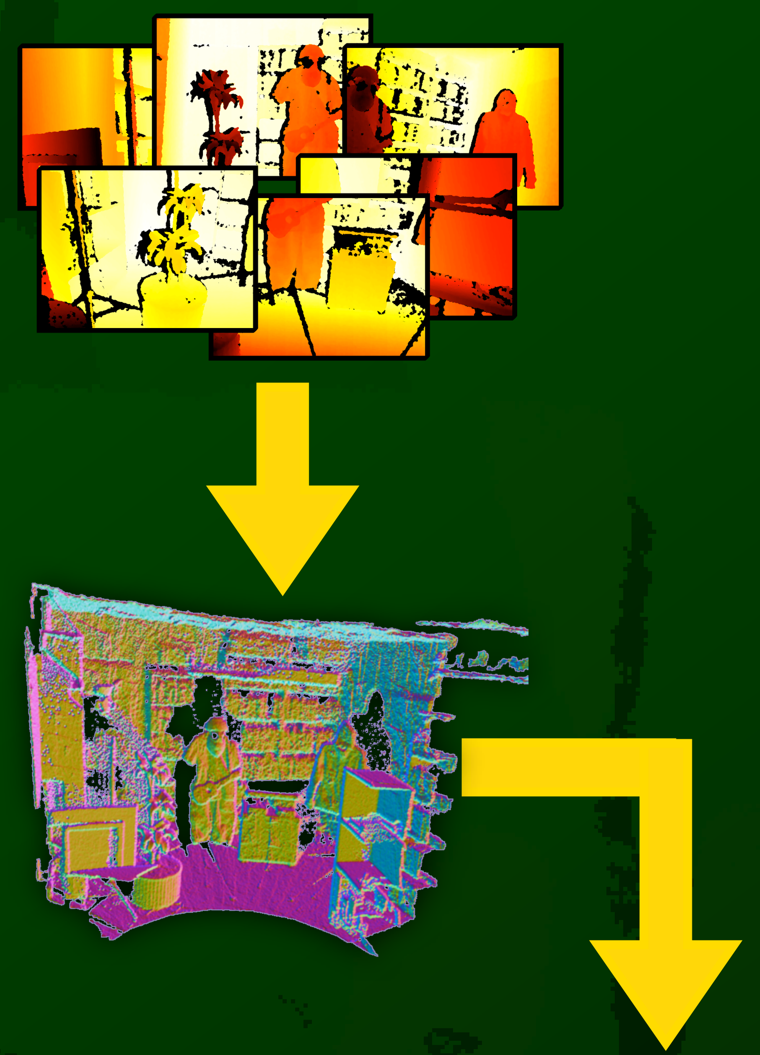
Pan-tilt stereo rig with Kinect. (A) - SLP projector, (B) - RGB camera, (C) - NIR camera, (D) - Left wide-angle camera, (E) - Right wide-angle camera, (F) - Diffusor (raised).

- Acquisition of wide-angle stereo images and ground truth is made using the pan-tilt rig shown on the left.
- PTU D46-17.5 pan-tilt unit.
- Point Grey Flea2 cameras with 2.5 mm wide-angle lenses (115° FoV).
- Microsoft Kinect.
- Calibration uses measurements of both inverse depth and pixel position.
  - Errors have different characteristics!
- Error variances in 2D and 3D measurements must be handled correctly.

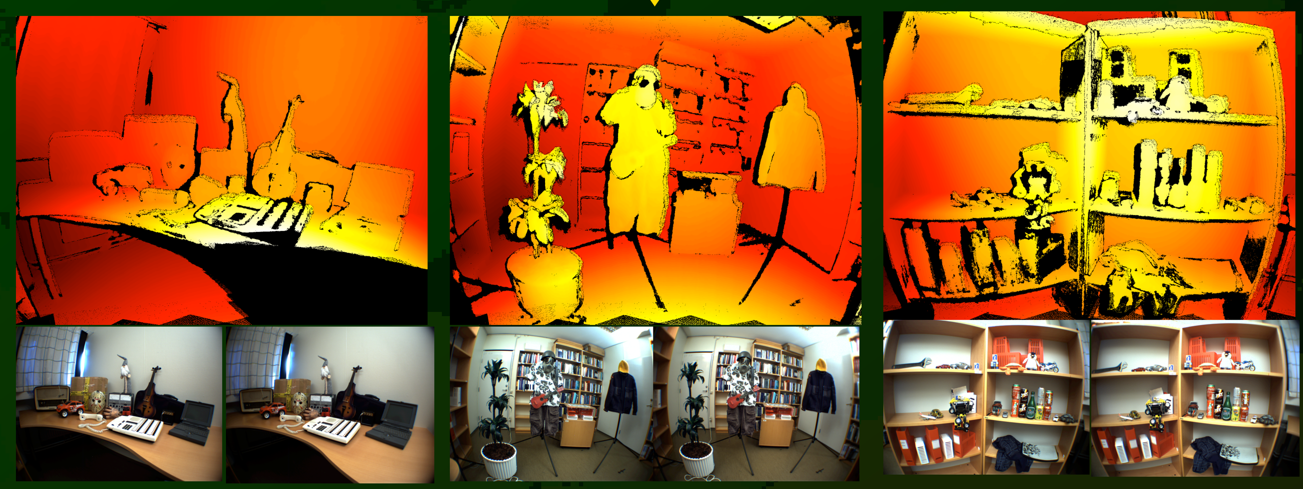


Propagate error variances between cameras and 3D points.

Ground-truth generation



- Wide-angle stereo ground-truth with high accuracy and low noise is required.
- The Kinect has a narrow field of view with limited accuracy in range. Range images contain noise.
- Many range scans from different angles are used to reconstruct 3D structure in a wide field of view.
- From the 3D structure, disparity is estimated at each pixel in the wide-angle image. A mean-shift algorithm is used to improve accuracy and reduce noise



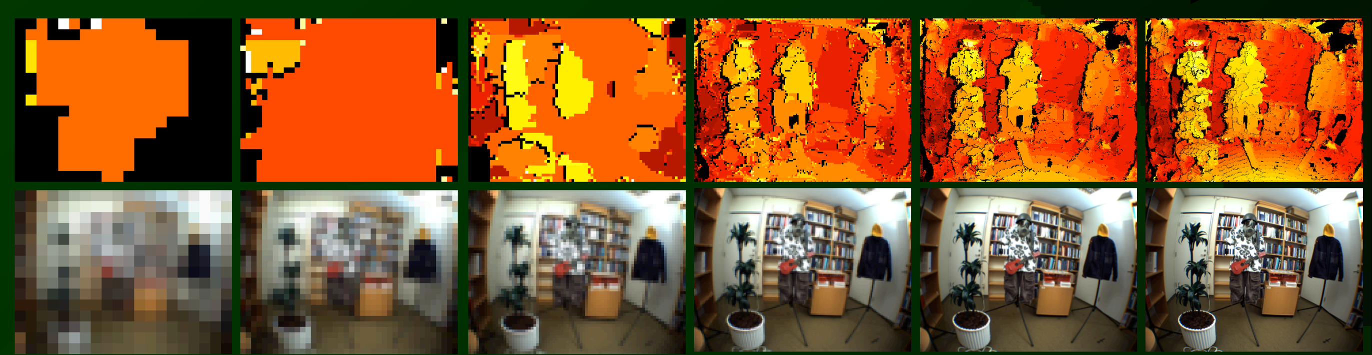
Examples of wide-angle ground-truth and stereo pairs from each of the three data sets.

Top row: disparity magnitude. Pixels not deemed reliably reconstructed are shown in black.

Bottom row: left and right images. Disparity map corresponds to left view.

Coarse-to-Fine BFP

- We extend the *best-first propagation* (BFP) algorithm with a coarse-to-fine scheme.
- From an initial assumption of identity mapping at coarse scale, correspondences are propagated in the image plane at progressively finer scales.
- At each scale, propagation is controlled by four parameters: window size, correlation threshold, structure threshold and sub-pixel refinement toggle.



Propagation of matches across multiple scales.

Top row, left to right: CtF-BFP results at progressively finer scales.

Bottom row, left to right: left camera image at corresponding scale.

Automatic tuning

- At each scale, CtF-BFP is controlled by four parameters (24 in total for six scales) → Automatic tuning is necessary!
- Optimisation must balance accuracy and robustness.
- Proposed objective function: $J(t_a, t_r) = \lambda r(t_r) - (1-\lambda)\int_0^{t_a} a(t)dt$

Rejection rate: $r(t_r) = \frac{1}{|\mathcal{V} \cap \mathcal{V}^*|} \sum_{(x,y) \in \mathcal{V} \cap \mathcal{V}^*} I(||\mathbf{D}^*(u,v) - \mathbf{D}(u,v)|| > t_r)$

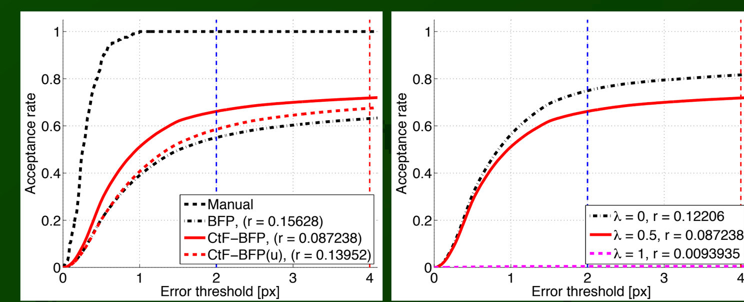
Acceptance rate: $a(t_a) = \frac{1}{|\mathcal{V}^*|} \sum_{(x,y) \in \mathcal{V} \cap \mathcal{V}^*} I(||\mathbf{D}^*(u,v) - \mathbf{D}(u,v)|| \leq t_a)$

The trade-off point between accuracy, coverage, and robustness is controlled by varying λ.

$\mathbf{D}$: Estimated disparities
$\mathcal{V}$: Valid estimated pixels
$\mathbf{D}^*$: Ground-truth disparities
$\mathcal{V}^*$: Valid ground-truth pixels

- Optimise each scale from coarse to fine
- Refine parameters from fine to coarse for best performance at finest scale.

Results

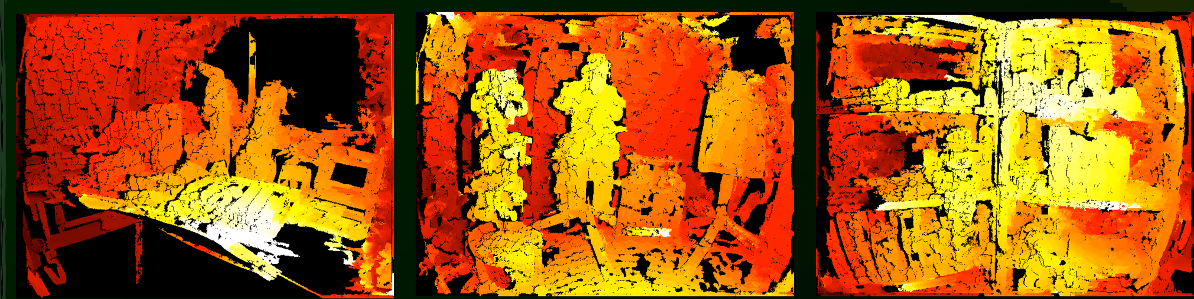


Left: Average acceptance curves over all data sets for automatically tuned CtF-BFP with λ = 0.5, BFP with original parameters, CtF- BFP(u) (before tuning). Errors on manually selected correspondences included as a best case.

Right: Average acceptance curves over all data sets for parameters tuned using λ = 0, 0.5, 1

- Results of tuning using one image from each data set, evaluated on different images (one from each data set).
- Cross-validation performed by tuning on five images from different poses in the same data set, evaluated on a sixth image from each data set.
  - Demonstrates balance between generalisation and adaptation.

$E \backslash T$	*Before tuning*	*Training set 1*	*Training set 2*	*Training set 3*
Evaluation set 1	-0.184	**-0.253**	-0.245	-0.240
Evaluation set 2	-0.054	-0.099	**-0.131**	-0.106
Evaluation set 3	-0.055	-0.116	-0.099	**-0.121**



Magnitude of disparity estimated using CtF-BFP for the example views from each data set.

The authors, and the embodied recognition platform Eddie are with the Computer Vision Laboratory, Linköping University.

Acknowledgements: This work was supported by the Swedish Research Council through a grant for the project *Embodied Visual Object Recognition*, and by Linköping University.

[1] : Marcus Wallenberg and Per-Erik Forssén.
Teaching Stereo Perception to YOUR Robot. In *BMVC*, 2012.
[2]: Maxime Lhuillier and Long Quan.
Match Propagation for Image-based Modelling and Rendering. In *IEEE TPAMI*, 24(8), 2002.