Teaching Stereo Perception to YOUR Robot

Marcus Wallenberg
http://users.isy.liu.se/cvl/wallenberg
Per-Erik Forssén
http://users.isy.liu.se/cvl/perfo

Computer Vision Laboratory
Linköping University
Linköping, Sweden

This paper describes a method for generation of dense stereo ground-truth using a consumer depth sensor such as the Microsoft Kinect. Such ground-truth allows adaptation of stereo algorithms to a specific setting. The method uses a novel residual weighting based on error propagation from image plane measurements to 3D. We use this ground-truth in wide-angle stereo learning by automatically tuning a novel extension of the best-first-propagation (BFP) dense correspondence algorithm. We extend BFP by adding a coarse-to-fine scheme, and a structure measure that limits its propagation along linear structures and flat areas. 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 such as BFP.

Wide-angle stereo provides an overview of a scene, even at very short range. The large *field of view* (FoV) also ensures that the visual fields from different points of view have a high degree of overlap. For these reasons, wide-angle lenses are popular in navigation, mapping and visual object search on robot platforms. However, the radial distortion caused by these lenses complicates the application of traditional stereo algorithms. A common approach to wide-angle stereo is to first attempt to remove radial distortion and then apply a descriptor-based wide-baseline stereo algorithm. An alternative approach is to use simpler matching metrics, and instead leverage *correspondence propagation*. One such algorithm is the *best-first propagation* (BFP) algorithm [2], and a recent addition is the *generalised PatchMatch algorithm* (GPM) [1]. Though these algorithms are more general than stereo algorithms, they have previously been applied to the stereo problem.

Since we make use of both the inverse depth and pixel coordinates in our calibration procedure, the effect of errors in these measurements must be taken into account during calibration. We therefore propagate error variances from these measurements into both 3D reconstructions and resulting 2D projections. We then fuse multiple Kinect range scans in a reference view coincident with the left stereo camera. This is done by estimating a disparity distribution for each pixel in this image, and using *mean-shift* to find the visible surface closest to the camera. We have imaged three indoor scenes, and for each of them calculated 51 full-resolution wide-angle disparity maps. Examples of individual range scans, an intermediate point cloud and the final disparity maps are shown in figure 2 (top row).

We use these to tune our extension of the BFP algorithm, which we call *coarse-to-fine best-first propagation* (CtF-BFP). Novelties are the use of multiple scales, a structure threshold that limits propagation along linear structures, and a sub-pixel refinement step. These novelties add a multitude of parameters, which make manual tuning difficult. We therefore use an automatic tuning procedure, that minimises an objective function that balances on accuracy, coverage and robustness.

When measuring performance of the stereo algorithm, we denote the estimated disparity map to be evaluated by $\mathbf{D}(u,v)$ defined on the domain $\mathcal{V}$ (pixels where disparities have been estimated). Similarly, the ground-truth disparity image is $\mathbf{D}^*(u,v)$, and the set of valid ground-truth pixels is $\mathcal{V}^*$.

To find a useful set of parameters, we minimise an objective function

$$J(t_a,t_r) = \lambda r(t_r) - (1-\lambda)\int_0^{t_a} a(t)dt. \qquad (1)$$

based on the acceptance rate $a(t_a)$ (the relative portion of disparities with an error below a threshold t_a) and rejection rate $r(t_r)$ (the relative portion of estimated pixels that differ from ground-truth by more than a threshold t_r). The parameter $\lambda \in [0,1]$ allows us to control the relative weight of the rejection and acceptance terms. Different values of λ produce different behaviour of $J(t_a,t_r)$ by influencing the trade-off between accuracy, coverage and robustness. Results of the automatic tuning procedure are shown in figure 3.

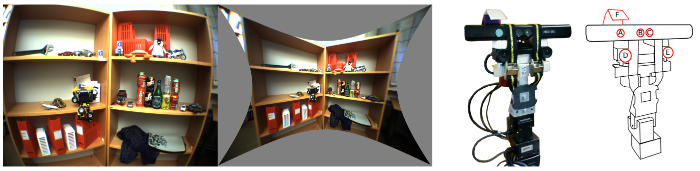

Figure 1: Left: In wide-angle images, angular resolution is near uniform. Middle: If they are rectified to preserve straight lines, most of the image is spent representing the periphery.
Right: Pan-tilt stereo rig with Kinect. (A) - SLP projector, (B) - RGB camera, (C) - NIR camera, (D) - Left wide-angle camera, (E) - Right wide-angle camera, (F) - Diffusor (raised).

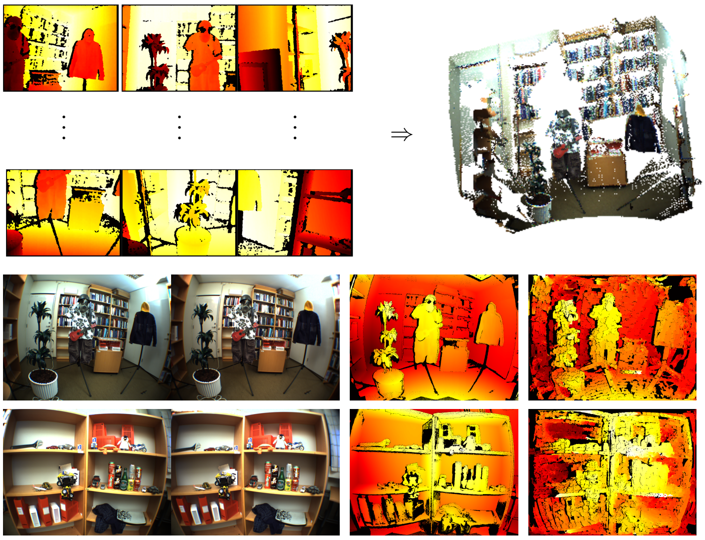

Figure 2: Top: Example of ground-truth generation. Left: Individual Kinect range scans. Right: several range scans projected into 3D.
Bottom: Examples of views for two scenes. Columns left to right: Left wide-angle image, right wide-angle image, magnitude of disparities deemed reliably reconstructed. The final column shows magnitude of disparity estimate obtained using tuned CtF-BFP.

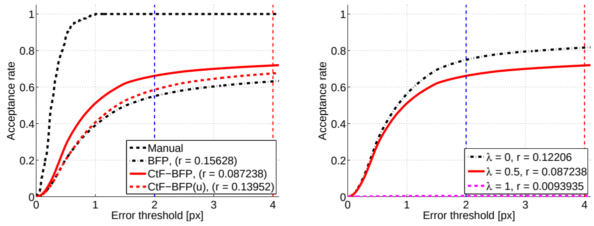

Figure 3: Left: Average acceptance curves over all data sets for automatically tuned CtF-BFP with $\lambda = 0.5$, BFP with original parameters, CtF-BFP(u) (before tuning). Errors on manually selected correspondences included as a best case. Right: Average acceptance curves over all data sets for parameters tuned using $\lambda = 0, 0.5, 1$.

[1] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized PatchMatch correspondence algorithm. In *European Conference on Computer Vision*, LNCS, September 2010.

[2] Maxime Lhuillier and Long Quan. Match propagation for image-based modelling and rendering. *IEEE TPAMI*, 24(8):1140–1146, 2002.