Combining Monocular Geometric Cues with Traditional Stereo Cues for Consumer Camera Stereo

Adarsh Kowdle, Andrew Gallagher, and Tsuhan Chen

Cornell University, Ithaca, NY, USA

Abstract. This paper presents an algorithm for considering both stereo cues and structural priors to obtain a geometrically representative depth map from a narrow baseline stereo pair. We use stereo pairs captured with a consumer stereo camera and observe that traditional depth estimation using stereo matching techniques encounters difficulties related to the narrow baseline relative to the depth of the scene. However, monocular geometric cues based on attributes such as lines and the horizon provide additional hints about the global structure that stereo matching misses. We merge both monocular and stereo matching features in a piecewise planar reconstruction framework that is initialized with a discrete inference step, and refined with a continuous optimization to encourage the intersections of hypothesized planes to coincide with observed image lines. We show through our results on stereo pairs of manmade structures captured outside of the lab that our algorithm exploits the advantages of both approaches to infer a better depth map of the scene.

Key words: narrow baseline stereo, consumer stereo camera

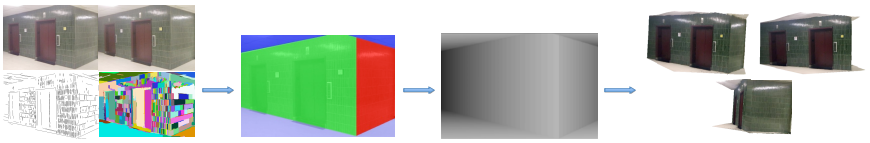
1 Introduction

Recent developments in consumer electronics have paved the way for handheld stereo cameras such as Fujifilm FinePix 3D® and Sony Bloggie 3D®, which allow users to capture stereo pairs in the wild (i.e. outside the lab). Using a *single* stereo pair captured with such a camera, we observe that while standard depth-from-stereo allows us to observe the ordering of various objects in the scene, even the state-of-art algorithms fail to give a good depth map that captures the geometry of the scene such as, the 3D structure of the facades of distant buildings and the depth of homogeneous surfaces (Fig. 3). The problems with stereo matching include: narrow baseline (typically 77 mm, similar to our eyes) that limits the depth from parallax [9], camera properties such as image resolution, and mismatches between sensors in terms of contrast, exposure, and focus, that lead to heavy distortion. In addition, problems due to scene irregularities such as ill-effects of lighting, specularities and homogeneous surfaces make the stereo matching task challenging.

However, humans (with eyes arranged similar to the cameras of a stereo camera), effectively use monocular cues from the scene to infer the 3D structure of the scene as illustrated in Fig. 1a. While the depth perception from stereo of the human visual system is also restricted to only a few meters, we have enough cues from monocular geometric cues and prior learning to obtain the global geometric structure of say, a building beyond the maximum depth from stereo. A number of recent works on depth from a single image have shown that one can obtain a fairly detailed depth map using trained models

(a) (b)

Fig. 1: (a) Monocular geometric cues: (Left) Müller-Lyer illusion: Equal line segments appear different since we tend to interpret them from a 3D geometry point of view; (Right) Prominent monocular geometric cues like vanishing points not captured in generic stereo matching reveal the structure of a scene even in the absence of texture cues; (b) Stereo cues: Ames room, a famous illusion gives you the above illusion when you look though a peephole (or a monocular image). Viewing it as a stereo pair, would rid of the illusion revealing that the girl on the left is far away compared to the girl on the right.



(a) (b) (c) (d)

Fig. 2: Overview: (a) Input stereo pair with the lines detected and superpixel map which help capture some of the monocular cues; (b) Resulting labeling using proposed algorithm; (c) Synthesized depth map (white is close, black is far); (d) 3D reconstruction from novel viewpoints.

[8,16,17,25]. Prior work on single-view 3D reasoning [7,22] and cognitive science have shown monocular *geometric* cues like lines and edges to be a critical aspect of human vision [3, 15]. While these help obtain a globally consistent structure, stereo cues can further disambiguate details or the depth ordering of objects in the scene as illustrated in Fig. 1b.

An overview of our approach and sample results are shown in Fig. 2 and Fig. 3. The main contributions of this paper are:

- We propose an algorithm to combine stereo cues with monocular structural priors (e.g. lines, horizon and plane intersections) that are not considered in generic stereo matching and prior works that combine monocular and stereo cues [24].
- We introduce monocular cues in two ways:
  1) By proposing possible parameterized planes. Stereo matching is used to then find the cost of assigning each plane to each superpixel. This is the discrete step.
  2) By continuous optimization that performs fine adjustment to encourage planes to meet at observed lines in the scene.
- We propose a novel use of distance transforms to encode monocular information from image lines within the discrete and continuous optimization steps.
- We show the effectiveness of the algorithm via a thorough comparison of the 3D reconstruction using a user study allowing users to fly-through the 3D rendering.

2 Related work

3D reconstruction of a scene is an active field in the community. Given a a single stereo pair of images, we give an overview of prior work to reconstruct the scene.

Depth from a single image. On one end of the spectrum is single-view modeling, which exploits solely monocular image cues [7, 8, 16, 17, 22, 25]. Some of these use

(a) (b) (c)

Fig. 3: Comparison with stereo matching (white is close, black is far for the depth maps): (a) Left image of stereo pair; (b) Depth map from stereo matching [29] shows errors such as the depth of the ceiling, and the depth of the piano; (c) Our result is geometrically representative of the scene.

trained models based on image features [16, 17, 25], under weak assumptions such as colinearity or coplanarity. A key idea that pervades single image reconstruction work is that some scene compositions (e.g. ground plane with perpendicular vertical planes) are more likely than others. We build on these by adding stereo analysis to our algorithm.

Multiview stereo. On the other end of the spectrum for 3D reconstruction is multiview stereo. A number of these approaches work with many images to obtain a fairly dense reconstruction [13, 27]. Recent works propose piecewise planar multiview stereo by using a discrete labeling over a set of hypothesized planes [11, 12, 14, 23, 28]. However, with a consumer stereo camera we face two problems: hypothesizing planes using a depth map is unreliable, and using multi-view stereo approaches using only photoconsistency across views does not perform well.

Depth from stereo. A natural approach to obtain depth from a stereo pair is to use dense stereo matching [1, 26]. While the stereo matching algorithms have been extensively evaluated on benchmarking datasets, we find that these algorithms are very sensitive to the data. Saxena et. al. [24] proposed making a depth map from traditional stereo matching. Holes in this map were filled using his learned single-view depth estimate. In practice, this tends to prefer a smooth reconstruction and given the poor performance of stereo matching on the consumer stereo pairs it would not recover geometric structures (intersecting planar surfaces).

In this work, we show that irrespective of inaccurate depth from stereo obtained due to scene irregularities, we can leverage the monocular geometric cues with the stereo cues to obtain a better, geometrically representative depth map.

3 Algorithm

In this section, we describe our algorithm in detail. We obtain the depth map of the scene in a two-step process. The first step is a discrete optimization that estimates a coarse structure of the scene, followed by a continuous optimization that refines the structure to render a geometrically representative depth map.

3.1 Discrete optimization

Motivated by recent works in piecewise planar stereo [12, 14, 23, 28] we try to achieve a globally consistent depth map of the scene by formulating depth estimation as a discrete optimization problem, where each pixel belongs to one of many hypothesized planes.

Plane hypothesis. We first calibrate the stereo camera to obtain the camera parameters (relative translation and rotation) of the two cameras. Given the camera parameters and matched SIFT features on the stereo pair, we estimate the 3D positions of the points resulting in a sparse point cloud. We hypothesize a set of dominant planes (L) by analyzing the distribution of depths of the 3D points along each hypothesized normals ([28]). Note that other plane hypotheses approaches can be used for this initial step [4, 14].

Energy minimization formulation. Let $i \in S$ denote superpixels in an image computed using color features from [10]. We describe an energy minimization formulation to estimate a labeling l, where each superpixel i is given a label $l_i \in L$. We define an MRF with the set of superpixels S as nodes, and all adjacent superpixels denoted as $\mathcal{N}$ as edges. We compute the labeling l that minimizes the following energy:

$$E(l) = \sum_{i \in \mathcal{S}} E_i(l_i) + \sum_{(i,j) \in \mathcal{N}} E_{i,j}\left(l_i, l_j\right), \tag{1}$$

where $E_i(l_i)$ is the unary term indicating the cost of assigning a superpixel i to a label l_i, and $E_{i,j}\left(l_i, l_j\right)$ is the pairwise term for penalizing label disagreement when neighboring superpixels i and j take the labels l_i and l_j, respectively.

Unary term Piecewise planar stereo algorithms typically capture this using a photoconsistency term measured over the multiple views. However, in case of a single narrow baseline stereo pair, a unary cost alone does not suffice. We model the term using monocular geometric cues in addition to the stereo photoconsistency term:

$$E_i(l_i) = \Psi(i, l_i) * \Big(E_i^P(l_i) + E_i^N(l_i)\Big), \tag{2}$$

where $E_i^P(l_i)$ is the photoconsistency term, $E_i^N(l_i)$ is a surface normal term and $\Psi(i, l_i)$ are additional monocular hard constraints we add to the unary term. We now explain these terms in detail.

Photoconsistency term (E^P)**.** The photoconsistency term is similar to that used in recent multiview stereo algorithms. For each superpixel on the left image (say), for every plane hypothesis we estimate the warp error (via homography) from the right to the left view, quantified using normalized cross correlation (NCC). We refer the reader to [28] for more details. We compute the NCC using the superpixel as support at each pixel as opposed to a constant window.

Surface normal term (E^N)**.** With a narrow baseline stereo pair, the two cameras lack enough parallax to differentiate planes beyond a certain depth. Monocular cues such as the lines and vanishing directions provide support to penalize the planes that would result in a globally inconsistent reconstruction.

We estimate the likelihood of each pixel belonging to a particular surface normal direction by developing a novel approach to exploit this information using the lines detected in the image. These lines were used to estimate the vanishing directions and hence hypothesize plane normals spanning the scene[1]. The lines are first assigned to one of the vanishing directions $\{\text{VD}\}$ as shown in Fig. 4. For each vanishing direction $v_p \in \{\text{VD}\}$, we compute a distance map $\left(\delta_{v_p}\right)$using all the lines assigned to v_p.

$$\delta_{v_p} = \min_{\textit{line} \in v_p} DT(\textit{line}), \tag{3}$$

[1] Vanishing points and horizon were computed using the algorithm by Kosecka et. al. [20]

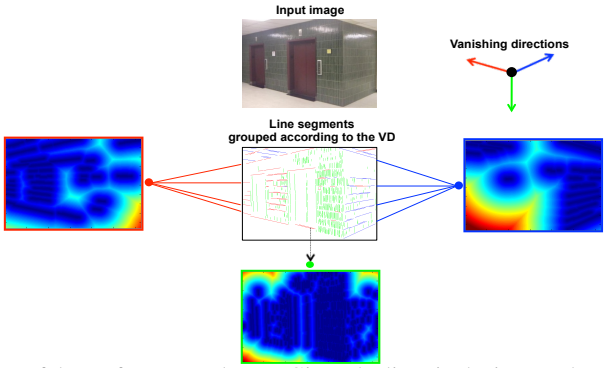


Fig. 4: Illustration of the surface normal term. Given the lines in the image, they are first grouped according to the vanishing directions (three in this case). The distance map for each vanishing direction is an ensemble of distance transforms with respect to the lines grouped to that direction. For example, the figure in the bottom shows the distance map for the green vanishing direction where, blue is a small distance transform value and red is large. Details in Section 3.1.

where, $DT(line)$ is the normalized distance transform, i.e., the distance of a pixel to the nearest point on the *line*. Now, considering the normal n_{l_i} of a plane l_i obtained using the cross product of two vanishing directions v_p and v_q. The per-pixel distance likelihood map for the surface normal n_{l_i} is estimated as:

$$D^{n_{l_i}} = \frac{(\delta_{v_p} + \delta_{v_q})}{2}, \tag{4}$$

The surface normal term represents the cost of superpixel i taking plane label l_i and is:

$$E_i^N(l_i) = \textit{median}\Big(D^{n_{l_i}}(p_i)\Big), \tag{5}$$

where p_i represents all pixels within superpixel i.

Monocular constraints (Ψ)**.** The monocular constraints for each superpixel i, $\Psi(i, l_i)$ gives the cost of choosing a plane that leads to an improbable scene. This depends on the normal of the hypothesized plane. We estimate the position of the horizon [17]. We add a large penalty for a superpixel above the horizon from choosing a plane with normal pointing upwards and for a superpixel below the horizon choosing a plane with normal downwards. For the space below the horizon, the penalty linearly decreases from 1.0 at the horizon to 0.0 at the bottom of the image. In case we fail to detect the horizon then this penalty is always 1.

Pairwise term We model the pairwise term using the well-known contrast-sensitive Potts model.

$$E_{i,j}(l_i, l_j) = \mathrm{I}(l_i \neq l_j)\ \exp(-\beta d_{ij}), \tag{6}$$

where $\mathrm{I}(\cdot)$ is an indicator function that is 1(0) if the input argument is true(false), d_{ij} is the contrast between superpixels i and j and β is a scale parameter.

Modeling the contrast term. (d_{ij}) The contrast is modeled as,

$$d_{ij} = O_{ij} * \left(\lambda_S\, d_{ij}^S + \lambda_C\, d_{ij}^C\right), \tag{7}$$

Fig. 5: Result of the discrete optimization. The discrete optimization stage assigns the correct normal, but it fails to distinguish between two parallel planes at different depths splitting the left wall into two regions (as shown by the two different shades of green) at different depths.

The first term (d_{ij}^S) is the stereo matching term. While the depth from stereo is not accurate, the precision is good to capture depth discontinuities. We model d_{ij}^S as $\frac{|d_i - d_j|}{d_j}$ where d_i and d_j are the mean disparities of the pixels within superpixel i and j respectively. The neighboring superpixels with disparity discontinuity are penalized if they take the same plane (and vice-versa). The second term (d_{ij}^C) is the normalized score from a coplanar classifier that captures the contrast between the features of adjacent superpixels [21]. If the coplanar classifier gives a high score for neighboring superpixels, it is penalized for not taking the same plane (and vice-versa). In order to handle inaccuracies in the contrast terms we obtain a soft normalized occlusion map [18]. We weight the pairwise term by the occlusion map, which intuitively captures the ambiguity of stereo matching algorithms in case of occlusions and allows label discontinuity across occlusions. The occlusion weight O_{ij} between superpixels i and j is given by the maximum occlusion confidence along the boundary between the two superpixels. λ_S and λ_C are regularization parameters that are manually tuned by observing the result on one stereo pair but are constant across all the datasets. In practice, equal weights gave good results on all stereo pairs.

Given the unary term and the pairwise term, we use graph-cuts with α-expansion to compute the MAP labels, using the implementation provided by Bagon [2] and Boykov et. al. [5, 6, 19]. The result on our sample stereo pair is shown in Fig. 5.

3.2 Continuous optimization

The discrete optimization quantizes the 3D space into meaningful planes that allow us to obtain a geometrically pleasing reconstruction. With a single stereo pair, the lack of parallax at large depths results in errors in the reconstruction; for example, it does not help distinguish between parallel planes at different depths (Fig. 5). We counter this problem by again using monocular cues, via a continuous optimization. This refinement stage will try to enforce the monocular constraint that we observe strong edges (or lines) when two non-coplanar surfaces meet.

Consider the erroneous region shown in Fig. 6b. We observe that the two adjacent planar regions say, π_m and π_n are segmented in the 2D fairly accurately. However, on projecting the 3D line of intersection onto the image, we observe that the plane estimate is inaccurate, Fig. 6c. We denote the projected line of intersection by the two points where the line meets the image *boundary* $(x_{mn,1}, y_{mn,1})$ and $(x_{mn,2}, y_{mn,2})$. We fix the plane normals, and using the 2D edge between the two segments we search the space of possible lines of intersection using the lines detected in the image. Once we obtain a target line of intersection defined by $(x'_{mn,1}, y'_{mn,1})$ and $(x'_{mn,2}, y'_{mn,2})$, (Fig. 6d) we optimize the plane parameters by minimizing the error function,

(a) (b) (c) (d) (e)

Fig. 6: Continuous optimization: (a) shows the result from discrete optimization; (b) highlights two regions that after the discrete step are labeled with the correct normals, but incorrect depths such that they are not connected (at the apparent intersection in the image); (c) shows the 2D projection of the line of intersection of planes represented by the highlighted regions (in red); (d) shows the target line of intersection which obeys image cues (in blue); (e) shows the final result, after the continuous optimization and refining the segmentation.

$$err(\pi_m, \pi_n) = |x_{mn,1} - x'_{mn,1}| + |y_{mn,1} - y'_{mn,1}| + |x_{mn,2} - x'_{mn,2}| + |y_{mn,2} - y'_{mn,2}| \quad (8)$$

The continuous optimization algorithm summarized in Algorithm 1. Fig. 6e shows the final result obtained by refining the segmentation using the new plane parameters.

Algorithm 1 Continuous optimization algorithm

1) Consider each region r_m where, $m \in \{1, 2, \ldots, N\}$ with plane parameters, $\pi_m = (\hat{n}_m, p_{(0,m)})$

2) Fix the normal $\hat{n}_m$ and optimize for $p_{(0,m)}$

3) Start from region r_i with highest number of non-parallel neighbors say $nei(r_i) = nei_1, nei_2, \ldots$

Do for each region r_i**:** {

Optimize for the vector of parameters,

$$p_0 = [p_{(0,i)}, p_{(0,nei_1)}, p_{(0,nei_2)}, \ldots]'$$

Constrained continuous optimization - bound the deviation of neighboring planes

$$\underset{p_0^*}{\operatorname{argmin}} \sum_{j \in nei(r_i)} err(\pi_i, \pi_j)$$

$$s.t : \forall j \neq i,\ p_{0,j} - \gamma < p_{0,j} < p_{0,j} + \gamma$$

where, γ decides the amount of deviation allowed for the neighboring planes.

}

4 Results and discussions

We perform our experiments on stereo pairs[2] captured using a recent consumer stereo camera, Fujifilm FinePix 3D W1®, which has a narrow baseline of 77mm. Given a stereo pair, we apply our algorithm to obtain plane parameters for each pixel. Given the 2D labeling and the plane parameters, we back-project to estimate the 3D position of each pixel. This allows us to synthesize the depth map, as well as render fly-throughs of the scene. We show the results of the algorithm in Fig. 7.

[2] `http://chenlab.ece.cornell.edu/projects/ConsumerCamStereo`

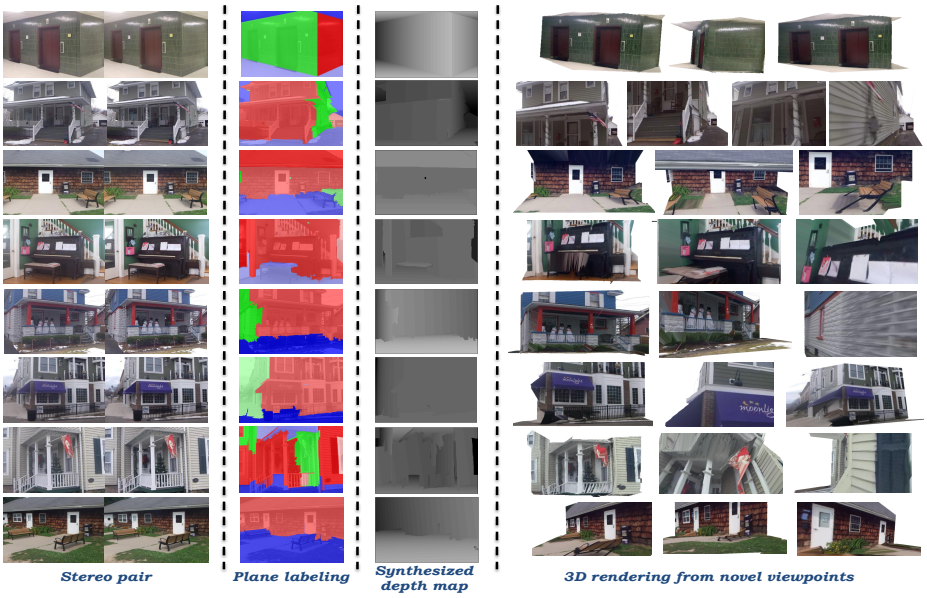


Fig. 7: Results (white=close and black=far for depth maps). Note that while the depth of the scene is more than 3 - 4 meters, given a *single* stereo pair of each scene we obtain depth maps that are geometrically representative. Row 2: while the stereo cues helped infer the porch being in front of the main facade, it is not strong enough (due to the depth of the scene) to infer the details of the porch. Row 4: the bench is correctly inferred as a horizontal region above the ground.

The importance of the continuous optimization stage can be shown with statistics. On an average, the algorithm resulted in eight unique plane labels for each scene in our dataset of ten stereo pairs. Each of these regions share support from an average of three non-parallel regions that contributes to refining the structure in the continuous optimization stage. The average error per region being optimized, computed using (8) decreased 72% from 87.1 to 24.0 pixels as a result of our continuous optimization stage.

4.1 Comparisons

We compare our work with other possible approaches to obtain the depth map of the scene with a single stereo pair.

We first compare the depth from stereo matching using [29] against our result in Fig. 3 and note that we perform better. Recent works have shown that we can obtain a reasonable depth map from a single image with prior trained models [8, 16, 17, 25]. We show some results in columns 1 and 2 of Fig. 8, and compare with our result. While Saxena et. al. enforce a smooth reconstruction without respecting the monocular geometry, Hoiem et. al. tend to rely on the ground segmentation, which results in inaccurate cutting and folding; we perform better than depth from a single image, which serves as a sanity check. While we do not re-implement work by Saxena et. al. [24], we note that due to the inaccurate depth map we would obtain a smooth reconstruction similar to column 1 in Fig. 8. Multi-view stereo approaches strongly rely on the photo-consistency constraint, and fails to differentiate between differently oriented planes as as shown in column 3 in Fig. 8. Micusik et. al. [23] encode some normal information using the spa-

tial structure of the superpixels but without the continuous optimization. Their results are less accurate, as shown in Fig. 5.

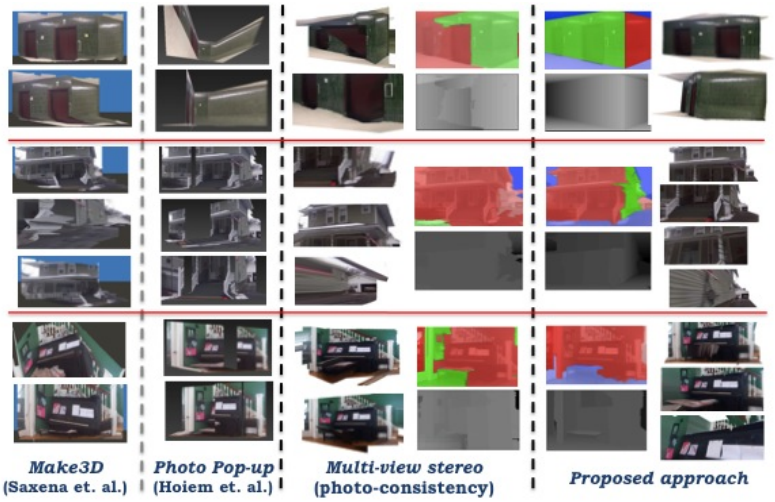


Fig. 8: Comparison with other approaches.

Qualitative comparison: User study We perform a qualitative comparison of our result, with single view modeling and multiview stereo via a psycho-visual study using 7 subjects. Each subject was presented results on ten stereo pairs without giving any indication about which of the three results they were looking at. They were given complete control to fly through the reconstructed scene and instructed to rank the three results from 1 (best) to 3 (worst) based on the geometric accuracy of the reconstruction. We expect the responses to be more consistent with relative ranking because absolute scores are hard to give, and need calibration across subjects. The average rank for single view modeling[3], multiview stereo and the proposed approach were obtained. The proposed approach was ranked as the best 69% of the time, more than triple the next best. This provides strong evidence that indicates the effectiveness of our approach.

5 Conclusions

We propose an algorithm to combine stereo cues with monocular structural priors to obtain geometrically accurate depth maps using stereo pairs captured in the wild using consumer stereo cameras. We introduce the idea of using both discrete and continuous optimization for 3D reasoning. Our approach leverages the use of monocular cues and exploits the benefits of discrete optimization to obtain a superpixel-to-plane labeling, followed by continuous optimization for refinement. We show through our results and comparisons that the proposed approach works well even in presence of homogeneous surfaces and specularities. The algorithm we propose can be used with any existing stereo matching algorithm, and additional monocular cues can easily be added to the same algorithm. For example, we can incorporate monocular cues such as depth from focus as a prior over the depth of different regions of the scene.

[3] The subject used the better result between [17] and [25] for ranking.

References

1. Middlebury stereo vision, *vision.middlebury.edu/stereo/*.
2. S. Bagon. Matlab wrapper for graph cut, December 2006.
3. H. G. Barrow and J. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17(1-3):75–116, 1981.
4. M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo - stereo matching with slanted support windows. In *BMVC*, 2011.
5. Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–1137, 2004.
6. Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *PAMI*, 20(12):1222–1239, 2001.
7. A. Criminisi, I. D. Reid, and A. Zisserman. Single view metrology. In *ICCV*, 1999.
8. E. Delage, H. Lee, and A. Y. Ng. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In *ISRR*, 2005.
9. J. Delon and B. Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223, 2007.
10. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.
11. Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *ICCV*, 2009.
12. Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Manhattan-world stereo. In *CVPR*, 2009.
13. Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *PAMI*, 2009.
14. D. Gallup, J.-M. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *CVPR*, 2010.
15. J. J. Gibson. Perception of the visual world. In *Houghton Mifflin*, 1950.
16. A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010.
17. D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*, 2005.
18. D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008.
19. V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–159, 2004.
20. J. Kosecká and W. Zhang. Video compass. In *ECCV*, 2002.
21. A. Kowdle, Y. Chang, A. Gallagher, and T. Chen. Active learning for piecewise planar multiview stereo. *CVPR*, 2011.
22. D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *CVPR*, 2009.
23. B. Micusík and J. Kosecká. Multi-view superpixel stereo in urban environments. *IJCV*, 89(1):106–119, 2010.
24. A. Saxena, J. Schulte, and A. Y. Ng. Depth estimation using monocular and stereo cues. In *IJCAI*, 2007.
25. A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *PAMI*, 31(5):824–840, 2009.
26. D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *SMBV*, 2001.
27. S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, 2006.
28. S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *ICCV*, 2009.
29. Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister. Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. *PAMI*, 31(3):492–504, 2009.