Make3D: Depth Perception from a Single Still Image

Ashutosh Saxena, Min Sun and **Andrew Y. Ng**
Computer Science Department,
Stanford University, Stanford, CA 94305
{asaxena,aliensun,ang}@cs.stanford.edu

Abstract

Humans have an amazing ability to perceive depth from a single still image; however, it remains a challenging problem for current computer vision systems. In this paper, we will present algorithms for estimating depth from a single still image.

There are numerous monocular cues—such as texture variations and gradients, defocus, color/haze, etc.—that can be used for depth perception. Taking a supervised learning approach to this problem, in which we begin by collecting a training set of single images and their corresponding ground-truth depths, we learn the mapping from image features to the depths. We then apply these ideas to create 3-d models that are visually-pleasing as well as quantitatively accurate from individual images. We also discuss applications of our depth perception algorithm in robotic navigation, in improving the performance of stereovision, and in creating large-scale 3-d models given only a small number of images.

Introduction

Upon seeing an image such as Fig. 1a, a human has no difficulty understanding its 3-d structure (Fig. 1c,d). However, inferring such 3-d structure remains extremely challenging for current computer vision systems. Indeed, in a narrow mathematical sense, it is impossible to recover 3-d depth from a single image, since we can never know if it is a picture of a painting (in which case the depth is flat) or if it is a picture of an actual 3-d environment. Yet in practice people perceive depth remarkably well given just one image; we would like our computers to have a similar sense of depths in a scene.

We view depth estimation as a small but crucial step towards the larger goal of image understanding, in that it will help in tasks such as understanding the spatial layout of a scene, finding walkable areas in a scene, detecting objects, etc. Most prior work on depth estimation has focused on methods that require multiple images, such as stereovision. These algorithms consider only the stereo (triangulation) cues (see related work section) and do not apply when only a single image is available. Beyond stereo/triangulation cues, there are also numerous *monocular* cues—such as texture variations and gradients, defocus, color/haze, etc.—that can be used to obtain rich 3-d information. In our approach, we capture some of these monocular cues using a Markov Random Fields (MRF). We take a supervised learning approach to this problem in which we use a 3-d laser scanner to collect training data comprised of a large number of images and their corresponding ground-truth depths. Using this training set, we learn the mapping from the image features to the depths. Our model also takes into account various other properties of the images, such as two adjacent regions in the image are more likely to be at the same depth, or to be even co-planar. Other than assuming that the environment is "locally planar," our model makes no explicit assumptions about the structure of the scene; this enables the algorithm to generalize well and to capture detailed 3-d structure.

In this paper, we will first discuss some of the visual cues that humans use for depth perception and describe how we construct similar features for using in our learning algorithm (Saxena, Chung, and Ng 2005; 2007). Then we will discuss how we used those ideas to produce visually-pleasing 3-d models from an image (Saxena, Sun, and Ng 2007b), and then describe an application of our depth perception algorithms to robotic navigation (Michels, Saxena, and Ng 2005). We will further describe how we use these ideas in improving the performance of stereo vision (Saxena, Schulte, and Ng 2007), and in creating large-scale 3-d models from a few images (Saxena, Sun, and Ng 2007a).

Algorithm

Humans use numerous visual cues to perceive depth. Such cues are typically grouped into four distinct categories: monocular, stereo, motion parallax, and focus cues (Loomis 2001). Humans combine these cues to understand the 3-d structure of the world (Wu, Ooi, and He 2004). Our probabilistic model attempts to capture a number of monocular cues, as well as stereo cues (when multiple images are available).

The monocular cues include texture variations, texture gradients, interposition, occlusion, known object sizes, light and shading, haze, defocus, etc. (Bülthoff, Bülthoff, and Sinha 1998) For example, the texture of surfaces appears different when viewed at different distances or orientations. A tiled floor with parallel lines will also appear to have tilted lines in an image, such that distant regions will have larger

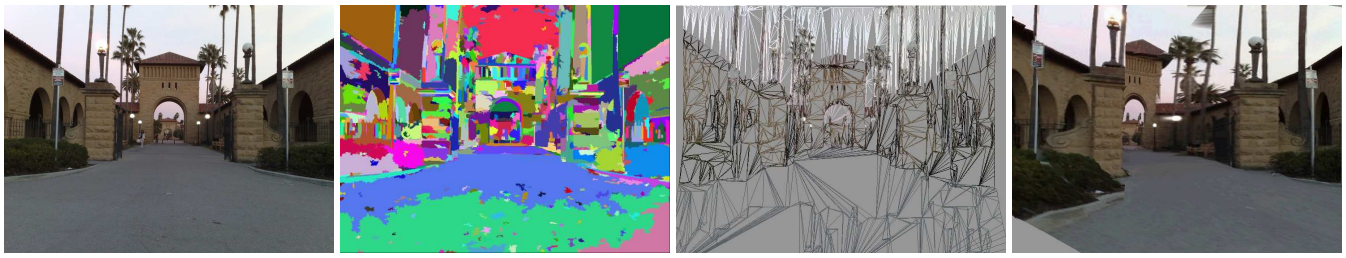

Figure 1: (a) An original image. (b) Oversegmentation of the image to obtain "superpixels". (c) The 3-d model predicted by the algorithm. (d) A screenshot of the textured 3-d model.

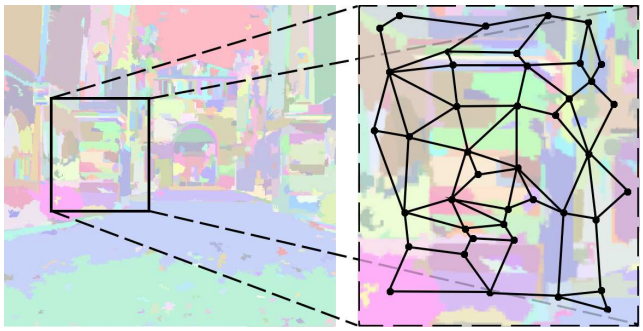

Figure 2: (Left) Superpixels in an image. (Right) An illustration of our MRF shown in a zoomed-in view. Each node corresponds to a superpixel in the image, and represents the 3-d position and 3-d orientation of the surface the superpixel came from; the edges represent the pairwise relations between two neighboring nodes.

variations in the line orientations, and nearby regions will have smaller variations in line orientations. Similarly, a grass field when viewed at different orientations/distances will appear different. An object will be smaller in the image if it is further away, and larger if it is closer.

Many monocular cues are "contextual information," in the sense that they are global properties of an image and cannot be inferred from small image regions. For example, occlusion cannot be determined if we look at just a small portion of an occluded object. Although local information such as the texture and color of a patch can give some information about its depth, this is usually insufficient to accurately determine its absolute depth. If we take a light blue patch, it is difficult to tell if it is infinitely far away (sky), or if it is a blue object. Due to ambiguities like these, one needs to look at the *overall* organization of the image to determine depths.

Images are formed by a projection of the 3-d scene onto two dimensions. Thus, given only a single image, the true 3-d structure is ambiguous, in that an image might represent an infinite number of 3-d structures. However, not all of these possible 3-d structures are equally likely. The environment we live in is reasonably structured, and thus humans are usually able to infer a (nearly) correct 3-d structure, using prior experience. In our learning algorithm, we try to capture the following properties of the images:

- **Image Features and depth**: The image features (textures, object sizes, etc.) bear some relation to the depth (and orientation) of a patch.
- **Connectivity**: Except in case of occlusion, neighboring patches are more likely to be connected to each other.
- **Co-planarity**: Neighboring patches are more likely to belong to the same plane, if they have similar features and if there are no edges between them.
- **Co-linearity**: Long straight lines in the image are more likely to be straight lines in the 3-d model: edges of a building, a sidewalk, a window, and so on.

Note that no single one of these four properties is enough, by itself, to predict the 3-d structure. Thus, our approach will combine these properties in a MRF (Fig. 2) in a way that depends on our "confidence" in each of these properties. Here, the "confidence" is itself estimated from local image cues.

In detail, our algorithm is as follows. We use the insight that most 3-d scenes can be segmented into many small, approximately planar surfaces. (Indeed, modern computer graphics using OpenGL or DirectX models extremely complex scenes this way, using triangular facets to model even very complex shapes.) Our algorithm begins by taking an image and using a segmentation algorithm (Felzenszwalb and Huttenlocher 2004) to find an oversegmentation of the image that divides it into many small regions (superpixels). An example of such a segmentation is shown in Fig. 1b. Because we use an over-segmentation, planar surfaces in the world may be broken up into many superpixels; however, each superpixel is likely to (at least approximately) lie entirely in only one planar surface.

For each superpixel, our algorithm then tries to infer the 3-d position and orientation of the 3-d surface it came from. This 3-d surface is not restricted to just vertical and horizontal directions, but can be oriented in any direction. The algorithm also infers the meaningful boundaries—occlusion boundaries or folds—in the image. Simply using an edge detector that relies just on local image gradients would be less reliable, because strong image gradients do not necessarily correspond to the occlusion boundary/fold, e.g., a shadow falling on a road may create an edge between the part with a shadow and the one without. Therefore, we use (supervised) learning to combine a number of such visual features to make the inference of such boundaries more accurate. Note that since our MRF "integrates" information from multiple cues, it would often be able to predict "correct" 3-d models even if the inference of these boundaries

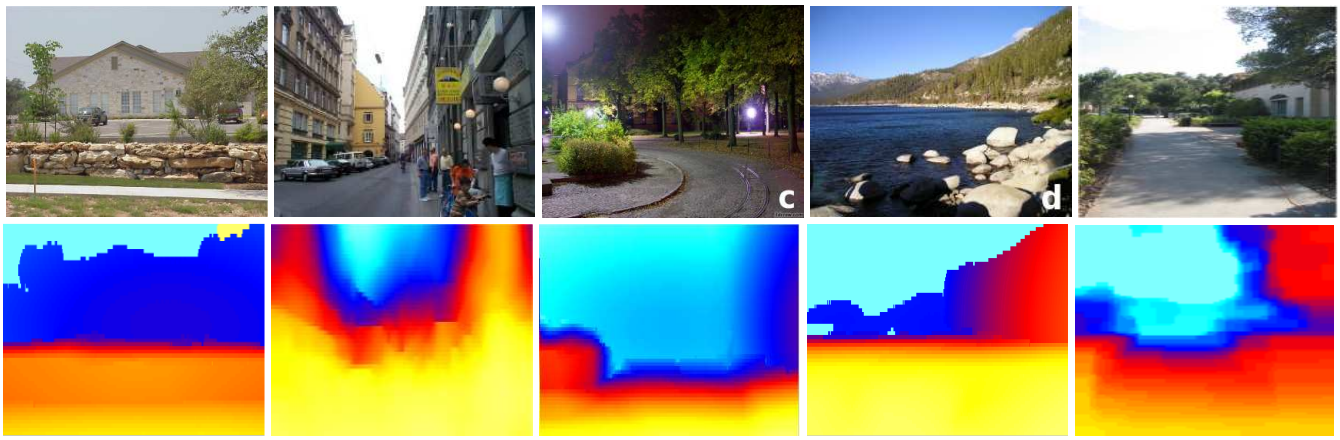

Figure 3: Typical results from our algorithm. (Top row) Original images downloaded from internet, (Bottom row) depths (shown in log scale, yellow is closest, followed by red and then blue) generated using our algorithm. (*Best viewed in color.*)

was not completely accurate.

Having inferred the 3-d position and 3-d orientation of each superpixel, we can now build a 3-d mesh model of a scene (Fig. 1c). We then texture-map the original image onto it to build a textured 3-d model (Fig. 1d) that we can fly through and view from different angles.

Results

We first applied our algorithm to the task of predicting depthmaps (i.e., depth at each point in the image) from a single image. In a simplified version of the algorithm, described in (Saxena, Chung, and Ng 2005; 2007), we used a point-wise representation of the image. In other words, instead of inferring the 3-d location and 3-d orientation of a superpixel, we inferred the 3-d location of each point in a uniform rectangular grid in the image. In a quantitative evaluation (against ground-truth depths collected using a laser scanner) on a test dataset of 107 images (data available online), we showed that our algorithm gave an error of 0.132 orders of magnitude, which corresponds to a multiplicative error of 35.5%. See Fig. 3 for some examples of the predicted depthmaps.

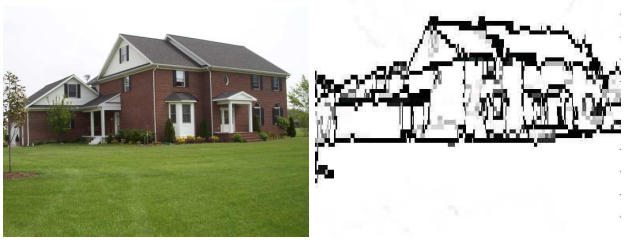

Figure 4: (Left) An image of a scene. (Right) Inferred "soft" values of the learned occlusion boundaries/folds.

However, to be able to render a textured 3-d model in a way that is also visually pleasing in addition to being quantitatively correct, we used our full model described in more detail in (Saxena, Sun, and Ng 2007b). We compared the following algorithms (see Table 1):

(a) Baseline: The Baseline MRF is trained without any image features, and thus reflects a "prior" depthmap of sorts.
(b) Our MRF: both the full model, as well as one without any priors.
(c) Hoiem et al. (HEH). For fairness, we scale and shift their depthmaps before computing the errors to match the global scale of our test images. Without the scaling and shifting, their error is much higher (7.533 for relative depth error).

We compare the algorithms on the following metrics: (a) % of models qualitatively correct, (b) % of major planes correctly identified.[1] (c) Average depth error on a log-10 scale, (d) Average relative depth error.

Table 1 first shows that our algorithm predicts quantitatively more accurate depths than both HEH and our first model (last subsection) (Saxena, Chung, and Ng 2005). In addition to producing quantitatively correct depths, our model does indeed produce significantly pleasing 3-d models. Our algorithm gives qualitatively correct models for 64.9% of images as compared to 33.1% by HEH. The qualitative evaluation was performed by a person not associated with the project. HEH generate a "photo-popup" effect by folding the images at "ground-vertical" boundaries—an assumption which is not true for a significant number of images; therefore, their method fails in those images. Further, our models are often more detailed than HEH, in that they are often able to model the scene with a multitude (over a hundred) of planes. (See Fig. 5 and Saxena et al., 2007b.) In another comparison: even when both HEH and our algorithm are evaluated as qualitatively correct on an image, one result could still be superior. Therefore, we asked the person to compare the two methods, and decide which one is better. We found that our algorithm performs better than HEH in 62.1% of the cases, while HEH performs better than us in 22.1% cases (tied in the rest).

In (Saxena, Sun, and Ng 2007b), we also described how to incorporate object recognition information into our model. For example, if we detect a standing person, we know that people usually stand on the floor and thus their feet must be

[1] For definition of "correct," see (Saxena, Sun, and Ng 2007b).

Table 1: Results: Quantitative evaluation.

METHOD	CORRECT (%)	% PLANES CORRECT	$\log_{10}$	REL
BASELINE	0%	0%	0.334	0.516
HEH	33.1%	50.3%	0.320	1.423
NO PRIORS	0%	0%	0.205	0.392
OUR MODEL	**64.9%**	**71.2%**	**0.187**	**0.370**

at ground-level. Knowing approximately how tall people are also helps us to infer their depth (distance) from the camera; for example, a person who is 50 pixels tall in the image is likely about twice as far as one who is 100 pixels tall. This allowed us to improve the accuracy of many models of images in which known objects were present.

Finally, in a large-scale web experiment, we allowed users to upload their photos on the internet and view a 3-d flythrough produced from their image by our algorithm. About 23846 unique users uploaded (and rated) 26228 images.[2] Users rated 48.1% of the models as good. If we consider the images of scenes only, i.e., exclude images such as company logos, cartoon characters, closeups of objects, etc., then this percentage was 57.3%. We have made the following website available for downloading datasets/code, and for conversion of an image to a 3-d model/flythrough:

http://make3d.stanford.edu

Such an online web service would (hopefully) find its use in various applications such as photo-sharing websites, and providing an immersive 3-d experience in the real-estate pictures, etc.

Figure 6: (Left) The remote-controlled car driven autonomously. (Right) A view from the car, with the chosen steering direction indicated by the red square; the estimated distances to obstacles in the different directions are shown by the bar graph below the image.

Robot Navigation

We used a simplified version of the algorithm for avoiding obstacles while autonomously driving a small car (Fig. 6a). In this problem, instead of predicting the depth at each pixel in the image, we only need to predict the depth of each column in the image, which corresponds to the distances to the obstacles in each possible steering direction. (Fig. 6b.) See (Michels, Saxena, and Ng 2005; Saxena, Chung, and Ng 2007) for more details.

We tested the algorithm by driving the car at four different locations, ranging from man-made environments with concrete tiles and trees to uneven ground in a forest environment with rocks, trees and bushes where the car is almost never further than 1m from the nearest obstacle. We used reinforcement learning to learn the control policy. The mean time before a crash ranged from 19 to more than 200 seconds, depending on the density of the obstacles. The unstructured testing sites were limited to areas where no training or development images were taken. Videos of the algorithm driving the car autonomously are available online at: **http://ai.stanford.edu/∼asaxena/rccar**

More recently, Plagemann et al. (2008) used the monocular depth predictions from an omni-directional camera as input to a mapping algorithm and produced large-scale 2-d maps of indoor environments.

Multiple Images

We believe that monocular cues and (purely geometric) stereo cues give largely orthogonal, and therefore complementary, types of information about depth. Therefore, we extended our model to combine both monocular and stereo cues—this gave better depth estimates than was possible with either alone (Saxena, Schulte, and Ng 2007; Saxena, Chung, and Ng 2007). Our algorithm estimated depths with an error of 0.074 orders of magnitude (i.e., 18.6% multiplicative error), which represented a significant improvement over stereovision's performance of 0.088. (See Fig. 7 for an example.)

Extending these ideas, we also considered the problem of creating 3-d models of large novel environments, given only a small, sparse, set of images. In this setting, some parts of the scene may be visible in multiple images, so that triangulation cues (structure from motion) can be used to help reconstruct them; but larger parts of the scene may be visible only in one image. We extended our model to simultaneously reason over multiple images and therefore build full, photo-realistic 3-d models of larger scenes. See (Saxena, Sun, and Ng 2007a) for a more detailed discussion of the algorithm and the results.

Related Work

Recovering 3-d depth from images is a basic problem in computer vision, and has important applications in robotics, scene understanding and 3-d reconstruction. Most prior work on visual 3-d reconstruction has focused on using methods such as stereovision (Scharstein and Szeliski 2002) and structure from motion (Forsyth and Ponce 2003), which require two (or more) images, and rely on triangulation to estimate depths. However, the algorithms relying on triangulation are fundamentally limited by the baseline distance between the two camera-positions—i.e., the depth estimates tend to be inaccurate when the distances considered are large. They also tend to fail for textureless regions where correspondences cannot be reliably found. Further, these

[2] No restrictions were placed on the type of images that users can upload. Users can rate the models as good (thumbs-up) or bad (thumbs-down).

Figure 5: Images downloaded from the internet (first and third row), and the snapshots of the 3-d flythroughs produced by our algorithm. (Notice the occlusions in the first, the fourth and the last image.)

methods do not apply when only a single image is available.

For a few specific settings, several authors have developed methods for depth estimation from a single image. Examples include shape-from-shading (Zhang et al. 1999; Maki, Watanabe, and Wiles 2002) and shape-from-texture (Malik and Rosenholtz 1997; Lindeberg and Garding 1993); however, these methods are difficult to apply to surfaces that do not have fairly uniform color and texture. Nagai et al. (2002) used Hidden Markov Models to performing surface reconstruction from a single image for known, fixed objects such as hands and faces. Hassner and Basri (2006) used an example-based approach to estimate depth of an object from a known object class. Han and Zhu (2003) performed 3-d reconstruction for known specific classes of objects placed in untextured areas. Criminisi, Reid and Zisserman (2000) provided an interactive method for computing 3-d geometry, where the user can specify the object segmentation, 3-d coordinates of some points, and reference height of an object. Torralba and Oliva (2002) studied the relationship between the Fourier spectrum of an image and its mean depth.

In work contemporary to this, Delage, Lee and Ng (2005; 2006) and Hoiem, Efros and Hebert (HEH) (2005b; 2005a) assumed that the environment is made of a flat ground with vertical walls. They produced "photo-popups" that made strong assumptions about the scene consisting of ground/horizontal planes and vertical walls (and possibly sky); these methods therefore do not apply to the many images that do not have this "ground-vertical" structure. For example, images of mountains (e.g., Fig. 1), trees, staircases (e.g., Fig. 5a), arches, rooftops (e.g., Fig. 5c) often have much richer 3-d structure.

Building on the idea of single image depth perception, Sudderth et al. (2006) and Saxena, Sun and Ng (2007b) used recognition of familiar objects for depth estimation, Hoiem, Efros and Hebert (2006) used knowledge of scene geometry to the task of object recognition, More recently, Nedovic et al. (2007) classified an image into a discrete number of classes, each representing a particular scene geometry.

Figure 7: (a) Image, (b) Ground-truth depthmap, (c) Stereo depthmap (the black color represents areas where depth values were missing, (d) Mono-only depthmap, (e) Stereo+mono depthmap.

Discussion

The problem of depth perception is fundamental to computer vision, one that has enjoyed the attention of many researchers and seen significant progress in the last few decades. However, the vast majority of this work, such as stereopsis, used multiple image geometric cues to infer depth. In contrast, single-image cues offer a largely orthogonal source of information, one that has heretofore been relatively underexploited. Given that depth and shape perception appears to be an important building block for many other applications, such as object recognition (Hoiem, Efros, and Hebert 2006), grasping (Saxena et al. 2006), navigation (Michels, Saxena, and Ng 2005), image compositing (Kawakita et al. 2004), and video retrieval (Ewerth, Schwalb, and Freisleben 2007), we believe that monocular depth perception has the potential to improve all of these applications, particularly in settings where only a single image of a scene is available.

References

Bülthoff, I.; Bülthoff, H.; and Sinha, P. 1998. Top-down influences on stereoscopic depth-perception. *Nature Neuroscience* 1:254–257.

Criminisi, A.; Reid, I.; and Zisserman, A. 2000. Single view metrology. *IJCV* 40:123–148.

Delage, E.; Lee, H.; and Ng, A. 2005. Automatic single-image 3d reconstructions of indoor manhattan world scenes. In *ISRR*.

Delage, E.; Lee, H.; and Ng, A. Y. 2006. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *CVPR*.

Ewerth, R.; Schwalb, M.; and Freisleben, B. 2007. Using depth features to retrieve monocular video shots. In *Int'l conf Image and Video retrieval*.

Felzenszwalb, P., and Huttenlocher, D. 2004. Efficient graph-based image segmentation. *IJCV* 59.

Forsyth, D. A., and Ponce, J. 2003. *Computer Vision : A Modern Approach*. Prentice Hall.

Han, F., and Zhu, S.-C. 2003. Bayesian reconstruction of 3d shapes and scenes from a single image. In *ICCV Wrkshp Higher-Level Knowl 3D Model Motion Anal.*

Hassner, T., and Basri, R. 2006. Example based 3d reconstruction from single 2d images. In *CVPR workshop on Beyond Patches*.

Hoiem, D.; Efros, A.; and Hebert, M. 2005a. Automatic photo pop-up. In *ACM SIGGRAPH*.

Hoiem, D.; Efros, A.; and Hebert, M. 2005b. Geometric context from a single image. In *ICCV*.

Hoiem, D.; Efros, A.; and Hebert, M. 2006. Putting objects in perspective. In *CVPR*.

Kawakita, M.; Iizuka, K.; Aida, T.; Kurita, T.; and Kikuchi, H. 2004. Real-time three-dimensional video image composition by depth information. In *IEICE Electronics Express*.

Lindeberg, T., and Garding, J. 1993. Shape from texture from a multi-scale perspective. In *ICCV*.

Loomis, J. 2001. Looking down is looking up. *Nature News and Views* 414:155–156.

Maki, A.; Watanabe, M.; and Wiles, C. 2002. Geotensity: Combining motion and lighting for 3d surface reconstruction. *IJCV* 48(2):75–90.

Malik, J., and Rosenholtz, R. 1997. Computing local surface orientation and shape from texture for curved surfaces. *IJCV* 23(2):149–168.

Michels, J.; Saxena, A.; and Ng, A. Y. 2005. High speed obstacle avoidance using monocular vision and reinforcement learning. In *ICML*.

Nagai, T.; Naruse, T.; Ikehara, M.; and Kurematsu, A. 2002. Hmm-based surface reconstruction from single images. In *ICIP*.

Nedovic, V.; Smeulders, A. W.; Redert, A.; and Geusebroek, J.-M. 2007. Depth information by stage classification. In *ICCV*.

Plagemann, C.; Endres, F.; Hess, J.; Stachniss, C.; and Burgard, W. 2008. Monocular range sensing: A non-parametric learning approach. In *ICRA*.

Saxena, A.; Driemeyer, J.; Kearns, J.; and Ng, A. 2006. Robotic grasping of novel objects. In *NIPS 19*.

Saxena, A.; Chung, S. H.; and Ng, A. Y. 2005. Learning depth from single monocular images. In *NIPS*.

Saxena, A.; Chung, S. H.; and Ng, A. Y. 2007. 3-D depth reconstruction from a single still image. *IJCV* 76(1).

Saxena, A.; Schulte, J.; and Ng, A. Y. 2007. Depth estimation using monocular and stereo cues. In *IJCAI*.

Saxena, A.; Sun, M.; and Ng, A. Y. 2007a. 3-D reconstruction from sparse views using monocular vision. In *ICCV workshop Virtual Representations & Modeling of Large-scale environments*.

Saxena, A.; Sun, M.; and Ng, A. Y. 2007b. Learning 3-d scene structure from a single still image. In *ICCV workshop 3D Representation and Recognition (3dRR-07)*.

Scharstein, D., and Szeliski, R. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV* 47.

Sudderth, E.; Torralba, A.; Freeman, W. T.; and Willsky, A. S. 2006. Depth from familiar objects: A hierarchical model for 3d scenes. In *CVPR*.

Torralba, A., and Oliva, A. 2002. Depth estimation from image structure. *PAMI* 24(9):1–13.

Wu, B.; Ooi, T.; and He, Z. 2004. Perceiving distance accurately by a directional process of integrating ground information. *Nature* 428:73–77.

Zhang, R.; Tsai, P.; Cryer, J.; and Shah, M. 1999. Shape from shading: A survey. *PAMI* 21:690–706.