3D reconstruction is not just a low-level task: retrospect and survey

Jianxiong Xiao
Massachusetts Institute of Technology
jxiao@mit.edu

Abstract

Although an image is a 2D array, we live in a 3D world. The desire to recover the 3D structure of the world from 2D images is the key that distinguished computer vision from the already existing field of image processing 50 years ago. For the past two decades, the dominant research focus for 3D reconstruction is in obtaining more accurate depth maps or 3D point clouds. However, even when a robot has a depth map, it still cannot manipulate an object, because there is no high-level representation of the 3D world. Essentially, 3D reconstruction is not just a low-level task. Obtaining a depth map to capture a distance at each pixel is analogous to inventing a digital camera to capture the color value at each pixel. The gap between low-level depth measurements and high-level shape understanding is just as large as the gap between pixel colors and high-level semantic perception. Moving forward, we would like to argue that we need a higher-level intelligence for 3D reconstruction. We would like to draw attention of the 3D reconstruction research community to put greater emphasis on mid-level and high-level 3D understanding, instead of exclusively focus on improving of low-level reconstruction accuracy, as is the current situation. In this report, we retrospect the history and analyze some recent efforts in the community, to argue that a new era to study 3D reconstruction at higher level is starting to come.

1. Introduction

Although an image is a 2D array, we live in a 3D world where scenes have volume, affordances, and are spatially arranged with objects occluding each other. The ability to reason about these 3D properties would be useful for tasks such as navigation and object manipulation. As humans, we perceive the three-dimensional structure of the world around us with apparent ease. But for computers, this has been shown to be a very difficult task, and have been studied for about 50 years in the 3D reconstruction community in computer vision, which has made significant progress. Especially, in the past two decades, the dominant research focus for 3D reconstruction is in obtaining more accurate depth maps [44, 45, 55] or 3D point clouds [47, 48, 58, 50]. We now have reliable techniques [47, 48] for accurately computing a partial 3D model of an environment from thousands of partially overlapping photographs (using keypoint matching and structure from motion). Given a large enough set of views of a particular object, we can create accurate dense 3D surface models (using stereo matching and surface fitting [44, 45, 55, 58, 50, 59]). In particular, using Microsoft Kinect (also Primesense and Asus Xtion), a reliable depth map can be obtained straightly out of box.

However, despite all of these advances, the dream of having a computer interpret an image at the same level as a two-year old (for example, counting all of the objects in a picture) remains elusive. Even when we have a depth map, we still cannot manipulate an object because there is no high-level representation of the 3D world. Essentially, we would like to argue that 3D reconstruction is not just a low-level task. Obtaining a depth map to capture a distance at each pixel is analogous to inventing a digital camera to capture the color value at each pixel. The gap between low-level depth measurements and high-level shape understanding is just as large as the gap between pixel colors and high-level semantic perception. Moving forward, we need a higher-level intelligence for 3D reconstruction.

This report aims to draw the attention of the 3D reconstruction research community to put greater emphasis on mid-level and high-level 3D understanding, instead of exclusively focus on improving of low-level reconstruction accuracy, as is the current situation. In Section 2, we retrospect the history to study the different views of paradigms in the filed that makes 3D reconstruction a complete low-level task, apart from the view at the very beginning of 3D reconstruction research. We highlighted the point that draws a clear difference for the field, and analyze the long-term implication for subconscious changing in the view for 3D reconstruction. In Section 3, we review the widely accepted "two-streams hypothesis" model of the neural processing of vision in human brain, in order to draw the link between computer and human vision system. This link allows us to conjecture recognition in computer vision to be the counterpart of ventral stream in human vision system, and reconstruction to be the counterpart of dorsal stream in human vision system. In Section 4, we provide brief survey on some recent efforts in the community that can be regarded as studies for 3D reconstruction beyond low level. Finally, We highlight some recent efforts to unify recognition and reconstruction, and argue that a new era to study 3D reconstruction at higher-level is starting to come.

2. History and Retrospect

Physics (radiometry, optics, and sensor design) and computer graphics study the forward models about how light reflects off objects' surfaces, is scattered by the atmosphere, refracted through camera lenses (or human eyes), and finally projected onto a 2D image plane. In computer vision, we are trying to do the inverse [49], *i.e.* to describe the world that we see in one or more images and to reconstruct its properties, such as shape. In fact, the desire to recover the three-dimensional structure of the world from images and to use this as a stepping stone towards full scene understanding is what distinguished computer vision from the already existing field of digital image processing 50 years ago.

Early attempts at 3D reconstruction involved extracting edges and then inferring the 3D structure of an object or a "blocks world" from the topological structure of the 2D lines [41]. Several line labeling algorithms were developed at that time [29, 8, 53, 42, 30, 39]. Following that, three-dimensional modeling of non-polyhedral objects was also being studied [5, 2], using generalized cylinders [1, 7, 35, 40, 26, 36] or geon [6].

Staring from late 70s, more quantitative approaches to 3D were starting to emerge, including the first of many feature-based stereo correspondence algorithms [13, 37, 21, 38], and simultaneously recovering 3D structure and camera motion [51, 52, 34], *i.e.* structure from motion. After three decades of active research, nowadays, we can achieve very good performance with high accuracy and robustness, for both stereo matching [44, 45, 55] and structure from motion [47, 48, 58, 50].

However, there is a significantly difference between these two groups of approaches. The first group represented by "block world", targets on high-level reconstruction of objects and scenes. The second group, *i.e.* stereo correspondence and structure from motion, targets on very low-level 3D reconstruction. For example, the introduction of structure from motion was inspired by "the remarkable fact that this interpretation requires neither familiarity with, nor recognition of, the viewed objects" from [52]. It was totally aware that this kind of 3D reconstruction at low level is just a milestone towards higher-level 3D understanding, and is not the end goal.

However, this message somehow got mostly lost in the course of developing better-performing system. In the past three decades, there are a lot more success we achieve for the low-level 3D reconstruction for stereo correspondence and structure from motion, than for the high level 3D understanding. For low-level 3D reconstruction, thanks to the better understanding of geometry, more realistic image features, more sophisticated optimization routine and faster computers, we can obtain a reliable depth map or 3D point cloud together with camera poses. In contrast, for higher-level 3D interpretation, because the line-based approaches hardly work for real images, this field diminished after a short burst. Nowadays, the research for 3D reconstruction almost exclusively focuses on only low-level reconstruction, in obtaining better accuracy and improving robustness for stereo matching and structure from motion. People seems to have forgetten the end goal of such low-level reconstruction, *i.e.* to reach a full interpretation of the scenes and objects. Given that we can obtain very good result on low-level reconstruction now, we would like to remind the community and draw attention to put greater emphasis on mid-level and high-level 3D understanding.

We should separate the approach and the task. The less success of line-based approach for high-level 3D understanding should only indicate that we need a better approach. It shouldn't mean that higher-level 3D understanding is not important and we can stop working on it. In another word, we should focus on designing better approaches for high-level 3D understanding, which is independent of the fact that line based approach is less successful than key-point and feature based approach.

3. Two-streams Hypothesis :: Computer Vision

In parallel to the computer vision researchers' effort to develop engineering solutions for recovering the three-dimensional shape of objects in imagery, perceptual psychologists have spent centuries trying to understand how the human visual system works. The two-streams hypothesis is a widely accepted and influential model of the neural processing of vision [14]. The hypothesis, given its most popular characterization in [17], argues that humans possess two distinct visual systems. As visual information exits the occipital lobe, it follows two main pathways, or "streams". The ventral stream (also known as the "what pathway") travels to the temporal lobe and is involved with object identification and recognition. The dorsal stream (or, "how pathway") terminates in the parietal lobe and is involved with processing the objects spatial location relevant to the viewer.

The two-streams hypothesis remarkably matched well with the two major branches of computer vision – recognition and reconstruction. The ventral stream is associated with object recognition and form representation, which is the major research topic for recognition in computer vision. On the other hand, the dorsal stream is proposed to be in-

Human vision	Computer vision	Low Level	Mid Level	High Level
Ventral stream	Recognition	Color value	Grouping & Alignment	Semantic → Context
Dorsal stream	Reconstruction	Distance value	Grouping & Alignment	Shape → Structure
Question to answer at each level		How to process signal?	Which are together?	What is where?

Table 1. Different levels and different streams for both human and computer vision systems.

volved in the guidance of actions and recognizing where objects are in space. Also known as the parietal stream, the "where" stream, this pathway seems to be a great counterpart of reconstruction in computer vision.

The two-steams hypothesis in human vision is the result of research for human brain. But the distinction of recognition and reconstruction in computer vision rise automatically from the researchers in the field without much awareness. The computer vision researchers naturally separate the vision task into such two major branches, based on the tasks to solve, at the computational theory level.

This interesting coincidence enables us to make further analysis of the research focuses in computer vision. For recognition, *i.e.* counterpart of ventral stream, it is widely accepted that the task can be divided into three levels, as shown in Table 1. However, there is not separation of the three levels for reconstruction, simply because the current research of reconstruction exclusively focus on the low level part only. The mid level and high level for 3D reconstruction are mostly ignored. A lot of researchers, especially the younger generations, are not aware of the existing of the problem.

Now, thanks to our analogy between human vision and computer vision, we can now try to answer what are the core tasks of three different levels of reconstruction. Since both ventral and dorsal stream start from the primary visual cortex (V1), we can expect that the low level task for reconstruction should be signal processing and basic feature extraction, such as V1-like features and convolution of Gabor-like filter bank, or time-sensitive filter bank for motion detection to infer the structure. The mid level focuses on grouping and alignment. By grouping, we mean the grouping of pixels within the current frame for either color of depth value, *i.e.* the segmentation of the image plane into meaningful areas. This can happen in both 2D and 3D [65, 54]. By alignment, we mean the matching of the current input with previous exposed visual experience, *e.g.* as matching of a local patch with patches in a training set [46]. The grouping happens within the current frame, and the alignment happens between the current frame and previous visual experience. In both cases, the fundamental computational task for this level is to answer "which are together?" For the high level of recognition, the task is to infer the semantic meaning, *i.e.* the categories of objects, and furthermore, the context of multiple objects in the scene. For the high level of reconstruction, the task is to recognize the shape of individual objects, and to understand the 3D structure of the scene, *i.e.* the spatial relationship of objects in the scene (a shape is on top of another shape). At the end of computation, together with both recognition and reconstruction, or ventral stream and dorsal stream, the vision system will produce answers for "what is where?"

4. 3D Beyond Low Level: A Modern Survey

As previously mentioned, after the "blocks world" line of works, the community almost exclusively focuses on low-level 3D reconstruction. Very recently, there is a new increasing attention on the higher-level 3D reconstruction. In this section, we briefly summarize some representative works towards this direction.

4.1. Pre-history: Single-view 3D Reconstruction

The dominant approach of two-view or multiple-view 3D reconstruction is on the low level reconstruction using local patch correspondences. The performance of such approach usually significantly outperforms other alternatives, such as reasoning about the lines, because parallax is the strongest cue in this situation. Therefore, there are very few works on higher-level reconstruction in this domain. However, for single view image as input, because there is no parallax between images to utilize, many approaches are forced to try to be smarter to reason about higher-level reconstruction task. Therefore, in this subsection, we will only focus on the 3D reconstruction on single-view images.

Pixel-wise 3D Reconstruction: There are a line of works on the reconstruction of pixel-wise 3D property. Using Manhattan world assumption, Coughlan and Yuille [9] proposed a Bayesian inference to predict the compass direction from a single image. Hoiem *et al.* [28] used local image feature to train classifier to predict the surface orientation for each patches. And Saxena *et al.* [43] also used local image feature to train classifier, but to infer the depth value directly, under a conditional random field framework.

Photo Pop-up: Beyond prediction of 3D property for local image regions, a slightly higher-level representation is to pop-up the photos. Hoiem *et al.* [27] built on top of [28] to use local geometric surface orientation to fit ground-line that separate the floor and objects in order to pop-up the vertical surface. This photo pop-up is not only useful for computer graphics application, but also introduce the notion of higher-level reconstruction, by grouping the lower level surface estimation output with regularization (*i.e.* line fitting). For indoor scenes, Delage *et al.* [11] proposed a dynamic Bayesian network model to infer the floor structure for autonomous 3D reconstruction from a single indoor image. Their model assumes a "floor-wall" geometry on the scene and is trained to recognize the floor-wall boundary in each column of the image.

Line-based Single View Reconstruction: There is also a nice line of works [20, 3, 4, 68] that focus on using lines to reconstruct 3D shapes for indoor images or outdoor buildings, mostly based on exploring the Manhattan world property of man-made environments. In particular, [20] designed several common rules for a grammar to parse an image combining both bottom-up and top-down information.

4.2. Beginning: 3D Beyond Low Level

The volumetric 3D reasoning of indoor layout marked the beginning of 3D reconstruction beyond low level. Yu *et al.* 2008 [66] inferred the 3D spatial layout from a single 2D image by grouping: edges are grouped into lines, quadrilaterals, and finally depth-ordered planes. Because it aimed to infer the layout of a room, it is forced to reason about the 3D structure beyond low level. Since then, several groups independently started working on 3D geometric reasoning. Lee *et al.* 2009 [33] proposed to recognize the three dimensional structure of the interior of a building by generating plausible interpretations of a scene from a collection of line segments automatically extracted from a single indoor image. Then, several physically valid structure hypotheses are proposed by geometric reasoning and are verified to find the best fitting model to line segments, which is then converted to a full 3D model. Beyond lines, Hedau *et al.* 2009 [22] made use of geometric surface prediction [28] to gain robustness to clutter by modeling the global room space with a parametric 3D "box" and by iteratively localizing clutter and refitting the box.

Going one step further, not only the room layout can be estimated, we also desire to estimate the objects in the clutter. Hedau *et al.* 2010 [23] showed that a geometric representation of an object occurring in indoor scenes, along with rich scene structure can be used to produce a detector for that object in a single image. Using perspective cues from the global scene geometry, they first developed a 3D based object detector. They used a probabilistic model that explicitly uses constraints imposed by spatial layout - the locations of walls and floor in the image - to refine the 3D object estimates.

To model the 3D interaction between objects and the spatial layout, Lee *et al.* 2010 [32] proposed a parametric representation of objects in 3D, which allows us to incorporate volumetric constraints of the physical world. They showed that augmenting current structured prediction techniques with volumetric reasoning significantly improves the performance.

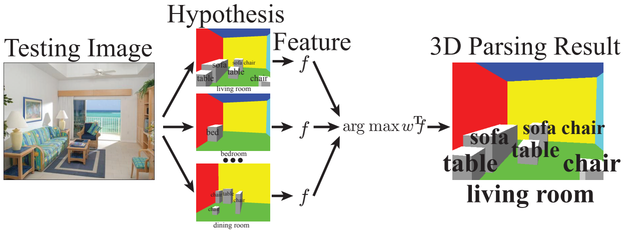

Figure 1. A system [64] that unifies recognition and reconstruction to recover semantic meaning and 3D structure at the same time.

On the other hand, going beyond indoor scenes, we can also reason about 3D structure for outdoor scenes. Also, previous approaches mostly operate either on the 2D image or using a surface-based representation, they do not allow reasoning about the physical constraints within the 3D scene. Gupta *et al.* [18] presented a qualitative physical representation of an outdoor scene where objects have volume and mass, and relationships describe 3D structure and mechanical configurations. This representation allows us to apply powerful global geometric constraints between 3D volumes as well as the laws of statics in a qualitative manner. They proposed an iterative "interpretation-by-synthesis" approach where, starting from an empty ground plane, the algorithm progressively "builds up" a physically plausible 3D interpretation of the image. Their approach automatically generates 3D parse graphs, which describe qualitative geometric and mechanical properties of objects and relationships between objects within an image.

Following these, there are many projects (*e.g.* [69, 10, 31, 24]) to reconstruct the 3D at higher level, especially at extraction of 3D spatial layout of indoor scenes, which becomes a very hot topic in major conferences of computer vision.

4.3. Unifying Recognition and Reconstruction

As illustrated in Table 1, eventually, we want the computer to answer "what is where?" for an image. Therefore, we have to combine the information from the output of recognition and reconstruction systems (or the ventral stream and dorsal stream in human vision). All the approaches mentioned above only focus on 3D reconstruction without any semantic meaning. We desire a system to do both: to predict the scene category [61, 57], the 3D boundary of the space, camera parameters, and all objects in the scene, represented by their 3D bounding boxes and categories. As shown in Figure 1, Xiao *et al.* [64] propose a unified framework for parsing an image to jointly infer geometry and semantic structure. Using a structural SVM, they

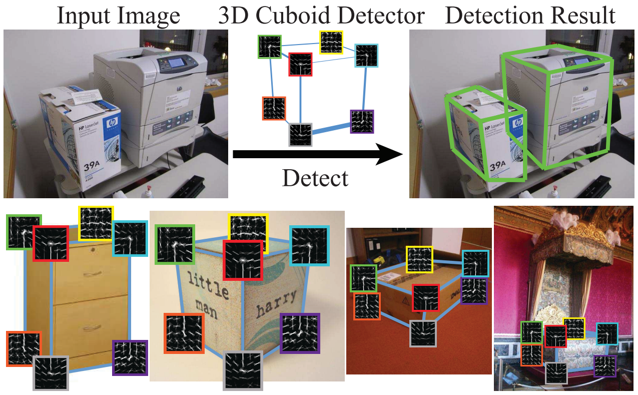

Figure 2. Shape Recognition: a 3D cuboid detector [63] that localize the corners of all cuboids in an image. This result will enable a robot to manipulate a cuboid-like object.

encode many novel image features and context rules into the structural SVM feature function, and automatically weigh the relative importance of all these rules based on training data. This demonstrates some initial results to jointly infer semantics and structures. For real applications, the unification of recognition and reconstruction can also be very useful, such as [62, 59, 56, 67].

4.4. Shape Recognition

Although higher-level 3D understanding starts with indoor room layout estimation, it is also very important for individual 3D object shape recognition. Very recently, there is a line of works that emphasis on this problem. In particular, for many objects, their 3D shape can be entirely explained by a simple geometric primitive, such as a cuboid. This is the case for most man-made structures [60, 59, 58]. Therefore, for such an image with cuboids, it would be very useful to parse the image to detect all the cuboids. Our desired output is not simply an indication of the presence of a geometric primitive and its 2D bounding box in the image as in traditional object detection. Instead, as shown in Figure 2, Xiao *et al*. [63] proposed a 3D object detector to detect rectangular cuboids and localize their corners in uncalibrated single-view images depicting everyday scenes. In contrast to the indoor layout based approaches that rely on detecting vanishing points of the scene and grouping line segments to form cuboids, they build a discriminative parts-based detector that models the appearance of the cuboid corners and internal edges while enforcing consistency to a 3D cuboid model. This model copes with different 3D viewpoints and aspect ratios and is able to detect cuboids across many different object categories.

Along the same line, [25, 15, 56] proposed 3D detectors for some object categories, such as cars and motorbikes. In particular, [25] proposed a two-stage model: the first stage reasons about 2D shape and appearance variation due to within-class variation (station wagons look different than sedans) and changes in viewpoint. Rather than using a view-based model, they described a compositional representation that models a large number of effective views and shapes using a small number of local view-based templates. They used this model to propose candidate detections and 2D estimates of shape. These estimates were then refined by their second stage, using an explicit 3D model of shape and viewpoint. They use a morphable model to capture 3D within-class variation, and use a weak-perspective camera model to capture viewpoint.

4.5. Human Activity for 3D Understanding

Human activity is a very strong cue for 3D understanding of scenes. Very recently, there is a line of works [19, 16, 12] pursuing this idea. [19] presented a human-centric paradigm for scene understanding. Their approach went beyond estimating 3D scene geometry and predicts the "workspace" of a human, which is represented by a data-driven vocabulary of human interactions. This method built upon the recent work in indoor scene understanding and the availability of motion capture data to create a joint space of human poses and scene geometry by modeling the physical interactions between the two. This joint space can then be used to predict potential human poses and joint locations from a single image.

On the other hand, [16] presented an approach which exploits the coupling between human actions and scene geometry. They investigated the use of human pose as a cue for single-view 3D scene understanding. Their method used still-image pose estimation to extract functional and geometric constraints about the scene. These constraints were then used to improve single-view 3D scene understanding approaches. They showed that observing people performing different actions can significantly improve estimates of 3D scene geometry.

5. Conclusion

While an image is a 2D array, we live in a 3D world. Although 3D reconstruction has been studied for nearly 50 years, recent progress in the field exclusively focus on very low-level reconstruction, such as recovering an accurate depth map or 3D point cloud. In this report, we argue that just like recognition, reconstruction is a task that contains all low-level, mid-level and high-level representation. We retrospect the history and analyze some recent efforts in the community, to argue that a new era to study 3D reconstruction at higher level is starting to come. We hope that this report draw attention of the 3D reconstruction research community to put greater emphasis on mid-level and high-level 3D understanding, instead of exclusively focus on improving of low-level reconstruction accuracy, to eventually build an intellegent vision machine.

Acknowledgement

This is the final report for the MIT 9.S912 "What is Intelligence?" class in Fall 2012. The instructors Prof. Shimon Ullman and Prof. Tomaso Poggio greatly inspired this work. Jianxiong Xiao is supported by Google U.S./Canada Ph.D. Fellowship in Computer Vision.

References

[1] G. Agin and T. Binford. Computer description of curved objects. *Computers, IEEE Transactions on*, 100(4):439–449, 1976.

[2] H. Baker. Three-dimensional modelling. In *Proceedings of the 5th international joint conference on Artificial intelligence*, pages 649–655, 1977.

[3] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin. Fast automatic single-view 3-d reconstruction of urban scenes. *Computer Vision–ECCV 2008*, pages 100–113, 2008.

[4] O. Barinova, V. Lempitsky, E. Tretiak, and P. Kohli. Geometric image parsing in man-made environments. *Computer Vision–ECCV 2010*, pages 57–70, 2010.

[5] B. Baumgart. Geometric modeling for computer vision. Technical report, DTIC Document, 1974.

[6] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[7] T. Binford, R. Brooks, and D. Lowe. Image understanding via geometric models. In *Proceedings of Fifth*, pages 364–9, 1980.

[8] M. Clowes. On seeing things. *Artificial intelligence*, 2(1):79–116, 1971.

[9] J. Coughlan and A. Yuille. Manhattan world: Compass direction from a single image by bayesian inference. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 941–947. IEEE, 1999.

[10] L. Del Pero, J. Guan, E. Brau, J. Schlecht, and K. Barnard. Sampling bedrooms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2009–2016. IEEE, 2011.

[11] E. Delage, H. Lee, and A. Ng. Automatic single-image 3d reconstructions of indoor manhattan world scenes. *Robotics Research*, pages 305–321, 2007.

[12] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros. Scene semantics from long-term observation of people. 2012.

[13] P. Dev. Segmentation processes in visual perception: A cooperative neural model. In *Proceedings of the 1974 Conference on Biologically Motivated Automata Theory, McLean, Va., June*, 1974.

[14] M. Eysenck. *Cognitive Psychology: A Student's Handbook: A Student's Handbook 5th Edition*. Psychology Press, 2005.

[15] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *Advances in Neural Information Processing Systems 25*, pages 620–628, 2012.

[16] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic. People watching: Human actions as a cue for single-view geometry. In *Proc. 12th European Conference on Computer Vision*, 2012.

[17] M. Goodale and A. Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[18] A. Gupta, A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. *Computer Vision–ECCV 2010*, pages 482–496, 2010.

[19] A. Gupta, S. Satkin, A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1961–1968. IEEE, 2011.

[20] F. Han and S. Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1778–1785. IEEE, 2005.

[21] P. Hans. Towards automatic visual obstacle avoidance. In *Proceedings of the 5th international joint conference on Artificial intelligence*, pages 584–584, 1977.

[22] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *Computer vision, 2009 IEEE 12th international conference on*, pages 1849–1856. IEEE, 2009.

[23] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. *Computer Vision–ECCV 2010*, pages 224–237, 2010.

[24] V. Hedau, D. Hoiem, and D. Forsyth. Recovering free space of indoor scenes from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2807–2814. IEEE, 2012.

[25] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In *Advances in Neural Information Processing Systems 25*, pages 602–610, 2012.

[26] G. Hinton. *Relaxation and its role in vision.* PhD thesis, University of Edinburgh, 1977.

[27] D. Hoiem, A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 577–584. ACM, 2005.

[28] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661. IEEE, 2005.

[29] D. Huffman. Impossible objects as nonsense sentences. *Machine intelligence*, 6(1):295–323, 1971.

[30] T. Kanade. A theory of origami world. *Artificial Intelligence*, 13(3):279–311, 1980.

[31] B. Kermgard. Bayesian geometric modeling of indoor scenes. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2719–2726, Washington, DC, USA, 2012. IEEE Computer Society.

[32] D. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating spatial layout of rooms using volumetric reasoning about ob-

jects and surfaces. *Advances in Neural Information Processing Systems (NIPS)*, 24:1288–1296, 2010.

[33] D. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2136–2143. IEEE, 2009.

[34] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, MA Fischler and O. Firschein, eds*, pages 61–62, 1987.

[35] D. Lowe and T. Binford. The interpretation of three-dimensional structure from image curves. In *Proceedings of IJCAI*, volume 7, pages 613–618, 1981.

[36] D. Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 1982.

[37] D. Marr and T. Poggio. Cooperative computation of stereo disparity. Technical report, DTIC Document, 1976.

[38] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 204(1156):301–328, 1979.

[39] V. Nalwa. *A guided tour of computer vision*. 1994.

[40] R. Nevatia and T. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8(1):77–98, 1977.

[41] L. Roberts. Machine perception of three-dimensional solids. Technical report, DTIC Document, 1963.

[42] A. Rosenfeld, R. Hummel, and S. Zucker. Scene labeling by relaxation operations. *Systems, Man and Cybernetics, IEEE Transactions on*, (6):420–433, 1976.

[43] A. Saxena, S. Chung, and A. Ng. Learning depth from single monocular images. *Advances in Neural Information Processing Systems*, 18:1161, 2006.

[44] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[45] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528. IEEE, 2006.

[46] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision*, 2012.

[47] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 835–846. ACM, 2006.

[48] N. Snavely, S. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.

[49] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition, 2010.

[50] P. Tan, T. Fang, J. Xiao, P. Zhao, and L. Quan. Single image tree modeling. *ACM Trans. Graph.*, 27(5):108:1–108:7, Dec. 2008.

[51] Ullman. *The Interpretation of Visual Motion*. Dissertation, MIT, 1977.

[52] S. Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979.

[53] D. Waltz. Generating semantic description from drawings of scenes with shadows. 1972.

[54] J. Xiao. Segmentation for image-based modeling. Final Year Thesis for Bachelor of Engineering in Computer Science, The Hong Kong University of Science and Technology, 2007. Undergraduate Thesis.

[55] J. Xiao, J. Chen, D.-Y. Yeung, and L. Quan. Learning two-view stereo matching. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 15–27, Berlin, Heidelberg, 2008. Springer-Verlag.

[56] J. Xiao, J. Chen, D.-Y. Yeung, and L. Quan. Structuring visual words in 3d for arbitrary-view object localization. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 725–737, Berlin, Heidelberg, 2008. Springer-Verlag.

[57] J. Xiao, K. Ehinger, A. Oliva, and A. Torralba. Recognizing scene viewpoint using panoramic place representation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2695 –2702, june 2012.

[58] J. Xiao, T. Fang, P. Tan, P. Zhao, E. Ofek, and L. Quan. Image-based façade modeling. *ACM Trans. Graph.*, 27(5):161:1–161:10, Dec. 2008.

[59] J. Xiao, T. Fang, P. Zhao, M. Lhuillier, and L. Quan. Image-based street-side city modeling. In *ACM SIGGRAPH Asia 2009 papers*, SIGGRAPH Asia '09, pages 114:1–114:12, New York, NY, USA, 2009. ACM.

[60] J. Xiao and Y. Furukawa. Reconstructing the world's museums. In *Proceedings of the 12th European conference on Computer Vision - Volume Part I*, ECCV'12, pages 668–681, Berlin, Heidelberg, 2012. Springer-Verlag.

[61] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3485 –3492, june 2010.

[62] J. Xiao and L. Quan. Multiple view semantic segmentation for street view images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 686 –693, 29 2009-oct. 2 2009.

[63] J. Xiao, B. Russell, and A. Torralba. Localizing 3d cuboids in single-view images. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 755–763. 2012.

[64] J. Xiao, B. C. Russell, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Basic level scene understanding: from labels to structure and beyond. In *SIGGRAPH Asia 2012 Technical Briefs*, SA '12, pages 36:1–36:4, New York, NY, USA, 2012. ACM.

[65] J. Xiao, J. Wang, P. Tan, and L. Quan. Joint affinity propagation for multiple view segmentation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –7, oct. 2007.

[66] S. Yu, H. Zhang, and J. Malik. Inferring spatial layout from a single image via depth-ordered grouping. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08.*

IEEE Computer Society Conference on, pages 1–7. IEEE, 2008.

[67] H. Zhang, J. Xiao, and L. Quan. Supervised label transfer for semantic segmentation of street scenes. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 561–574, Berlin, Heidelberg, 2010. Springer-Verlag.

[68] P. Zhao, T. Fang, J. Xiao, H. Zhang, Q. Zhao, and L. Quan. Rectilinear parsing of architecture in urban environment. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 342 –349, june 2010.

[69] Y. Zhao and S.-C. Zhu. Image parsing via stochastic scene grammar. In *Advances in Neural Information Processing Systems*, 2011.