What vision can, can't and should do

Michael Zillich

Vienna University of Technology

Abstract

Computer vision has come a long way since its beginnings. In this article we review some of the recent successes, which seem to indicate that many aspects of vision have indeed been solved and that the way should now be paved for robotic systems that can operate freely in the real world. On closer inspection though that is not the case just yet. A set of specialised solutions in different sub areas, however impressive individually, does not constitute a unified theory of vision. We point out some of the problems of current approaches, most notably lack of abstraction and dealing with uncertainty. Finally we suggest where research should and should not focus on in order to advance on a broader basis.

1. Introduction

Computer vision has made huge advances since its beginnings in the 1960's. After a slow start, plagued by limited computing power and sometimes overly optimistic predictions, recent years have seen increasing numbers of real world applications appearing on the market, from face tracking in consumer digital cameras, driving assistance systems in cars, autonomous vacuum cleaning robots to augmented reality applications or home entertainment. Of course industrial machine vision confined to the clearly structured environments of factory floors and assembly lines or medical imaging applications with a human in the loop have been on the market far longer. But within the scope of this article we are interested in computer vision as it was seen by its early proponents as exemplified by (Roberts, 1965; Binford, 1971; Clowes, 1971; Huffman, 1971; Waltz, 1975; Nevatia and Binford, 1977; Marr, 1982; Biederman, 1987): to understand the computational principles that allow human or animal vision to seemingly arrive at generic scene interpretations from images. Or put another way, vision that serves an agent to operate in and interact with the unconstrained real three-dimensional world.

This is of course a very broad definition and encompasses many different abilities related to locomation, manipulation, learning, recognition or social interactions. Part of the recent successes of vision, apart from increased computing power and the availability of new mathematical tools, is a high degree of specialisation of solutions in each of these areas. For many of these specialised solutions impressive videos can be watched online and one is left wonderering "Ok, done! So, where is the problem?". Yet

performance of robots at competitions like the Semantic Robot Vision Challenge [1] or RoboCup@Home [2], while clearly progressing from year to year, show that still these parts do not necessarily make a whole. There are of course many more problems to be solved in a complete robotic system besides vision, such as issues of power consumption and dexterity in manipulation, but limitations in percption and most notably vision typically do play a central role. So something must be still missing.

In the next section we will review a selection of state of the art solutions in some of the respective areas, showing which impressive things computer vision can in fact already do. Section 3 will then try to identify some fundamental problems in the current approach to computer vision, followed by suggestions on where future research could advance on a broader basis in Section 4.

2. What Vision Can Do

The following selection of work is not intended as a genuine review of work in different areas of computer vision, but rather to highlight some state of the art solutions that taken together could seem to have solved computer vision for robotics. So, what can vision do for robotics?

Navigation, Localisation. Self localisation and mapping (SLAM) has been addressed by the robotics community early on, starting with ultrasonic and later laser range sensors where it can essentially be considered solved (Thrun et al., 2005). With increasing computational power and mathematical tools such as sparse bundle adjustment (Lourakis and Argyros, 2009) vision based methods (Visual SLAM) began to replace laser based ones (e.g., Nistér et al., 2006; Davison et al., 2007), which can now handle very large areas (Cummins and Newman, 2010).

These methods rely on the robust extraction of uniquely identifyable image regions, for which a variety of image features have been proposed, such as MSER (Matas et al., 2002), SIFT (Lowe, 2004), FAST (E. Rosten and T. Drummond, 2006), SURF (Bay et al., 2008) or DAISY (Tola et al., 2008). These features in general play an essential role in many modern computer vision approaches, from SLAM and structure from motion to object recognition and tracking.

3D Reconstruction. Using similar techniques, structure from motion (SfM) approaches put an emphasis of dense reconstruction of the scene rather than navigation based on sparse landmarks. Microsoft's Photo Tourism (Snavely et al., 2006) is quite well known. It can reconstruct in very high detail a building such as the cathedral of Notre Dame from a colection of thousands of photographs taken from the web. Going even further (Agarwal et al., 2009) scales the approach to entire cities, although that does take, as the title of their paper suggests, a day of computation on a cluster of 500 computers.

Real time solutions are available for smaller scale scenes. The approach by (Klein and Murray, 2007) uses a parallel processing pipeline highly optimised to today's

[1] http://www.semantic-robot-vision-challenge.org
[2] http://www.ai.rug.nl/robocupathome

multi-core machines to build a semi-dense map of the environment based on tracking distinctive image features. Based on that the work by (Newcombe and Davison, 2010) fills in the details using GPU-based optical flow computation (Zach et al., 2007) to arrive at a dense 3D scene reconstruction with visually very pleasing results.

Scene Segmentation. The above approaches reconstruct the scene as a whole, essentially treating it as a single rigid and static object. Multibody structure from motion approaches (e.g., Fitzgibbon and Zisserman, 2000; Ozden et al., 2010) observe a dynamic scene and segment it into independently moving rigid objects.

Given only a static scene (Rusu et al., 2009) segments a 3D point cloud as provided by stereo or depth sensors into parametric object models such as planes, spheres, cylinders, and cones. Similarly (Biegelbauer et al., 2010) fit superquadrics to point clouds to seamlessly cover a wider range of parametric shapes. Using a strong prior model of the 3D scene and again parametric object models (Hager and Wegbreit, 2011) is able to handle scenes exhibiting complex support and occlusion relations between objects, and also reasons explicitly about dynamic changes of the scene such as objects being moved, added or removed.

Taking a more active approach (Björkman and Kragic, 2010) combine wide angle and foveated stereo to segment 3D objects of arbitrary shape standing isolated on a supporting surface. Even more active, (Fitzpatrick and Metta, 2003) use a robot manipulator to poke parts of the scene in order to use the resulting motion in 2D image sequences to segment objects.

Recognition. Object recognition is of course a central theme in computer vision especially in the context of robotics. Early attempts at generic recognition of 3D solids (e.g., Binford, 1971; Waltz, 1975; Nevatia and Binford, 1977; Marr and Nishihara, 1978; Brooks, 1983; Biederman, 1987; Lowe, 1987; Dickinson et al., 1992), often based on edge features, tended to suffer from scene complexity and textured surfaces. With the advent of invariant interest point detectors (Mikolajczyk and Schmid, 2004) and strongly distinctive point descriptors mentioned above (Matas et al., 2002; Lowe, 2004; E. Rosten and T. Drummond, 2006; Bay et al., 2008; Tola et al., 2008) appearance based recognition of arbitrarily shaped object instances in highly cluttered real world environments was essentially solved (e.g., Lowe, 1999; Gordon and Lowe, 2006; Ferrari et al., 2006; Özuysal et al., 2007; Collet et al., 2009; Mörwald et al., 2010), even for non-rigid objects such as clothing (Pilet et al., 2007) – provided of course that the respective objects are textured.

Recent advances in 3D sensing, most notably the Microsoft Kinect RBD-D sensor, brought a renewed interest in 3D methods. Making use a combination of color image and dense depth map the fast template based approach by (Hinterstoisser et al., 2011) also detects untextured objects in heavy clutter at close to frame rate.

The above appearance based methods are intrinsically suited to detect individual object instances with specific surface markings. Going beyond single instances approaches such as (Fei-Fei et al., 2006; Leibe and Schiele, 2003; Dalal and Triggs, 2005) detect categories also of deformable objects such as cows or walking humans.

Online Learning. Acquiring models for the above recognition methods often involves hand-labeling of images or placing objects on turn tables as an offline learning step, which is clearly not desirable for an agent supposed to act autonomously in the world.

Various online learning methods have been proposed, such as (Özuysal et al., 2006) which keeps "harvesting" additional features as it tracks the model acquired so far. The ProFORMA system (Pan et al., 2009) even reconstructs high quality dense triangle meshes while tracking a model and also suggesting new views to add.

Going further in the direction of a complete system (Kraft et al., 2008) and (Welke et al., 2010) let a robot pick up and rotate objects in its hand to actively cover all views of an object.

Tracking. Much as recognition, model based 3D object tracking has been well covered in computer vision (Lepetit and Fua, 2005). Especially with the availability of cheap and powerful graphics cards computationally heavy methods such as particle filtering (Klein and Murray, 2006) have been rendered real time (Chestnutt et al., 2007; Murphy-Chutorian and Trivedi, 2008; Sánchez et al., 2010; Choi and Christensen, 2010; Mörwald et al., 2011) and allow tracking of complex 3D objects through heavy clutter.

So far for an (incomplete) overview of some of the success stories of computer vision in the realm of robotics. Next we will look at where we stand with this and why service robots are not yet scurrying around in our appartments.

3. What Vision Can't Do

What vision can't do is simply to allow a robot to operate in and interact with the unconstrained real three-dimensional world, as was our stated goal in the introduction.

Abstraction. One of the reasons why this is the case is explored in the very comprehensive review by (Dickinson, 2009). The author there sums up the evolution of object categorization over the past four decades as different attempts to bridge the large representational gap between the raw input image at the lowest level of abstraction and 3D, viewpoint invariant, categorical shape models at the highest level of abstraction. In the 1970's this gap was closed by using idealised images of textureless objects in controlled lighting to extract quite generic shape models. In the 1980's the images could become more complex, by sacrificing model generality and searching for specific 3D shapes, thus effectively closing the gap at a lower level. Methods of the 1990's allowed recognition of complex textured objects in cluttered scenes, however objects were now esentially 2D appearance models of specific instances (and even views), thus closing the gap very low at the image level. The feature-based methods of the 2000's allowed recognition of arbitrary 3D object instances in very cluttered environments, while also slowly extending generality back up towards object categories.

So, much of the success of vision was bought by sacrificing generality and the ability of abstraction. This is less of a problem for navigation, where anything is an obstacle or a landmark, but more so for purposeful interaction with specific parts of the scene, viz, objects. Learning each object individually or perhaps narrow categories of objects is not feasible in the long run and leaves out the ability to make sense of

a scene even if a similar scene was never encountered. Humans, say an Inuit seeing tropical jungle for the first time, have no problem perceiving a completely unfamilier scene in terms of complete 3D shapes plus their possibilities of interaction rather than an assortment of object and category labels. Otherwise the mentioned Inuit would have to appear essentially cortically blind, having no categories for all the different tropical trees and bushes.

Putting it Together. Another reason as to why the assorted successes of vision do not yet comprise a unified solution for robotic vision lies in the difficulties of merging these specialised solutions under one framework. One of the difficulties is robustness. Many methods rely on tuning of parameters or some hidden implicit assumptions. Operating these methods outside their safe zones can make them fail abruptly rather than degrade gracefully, leaving a system comprised of many such isolated solutions extremely brittle. A more severe problem actually lies in bridging the semantic gaps between different methods. What does it mean if an object recogniser reports a confidence of 0.4 of detecting an object right inside a wall while robot localisation reports an uncertainty of 40 *cm*? Approaches like (Hoiem et al., 2006) have started exploring the interplay between e.g., object recognition and estimation of coarse 3D scene geometry, and the work by (Hager and Wegbreit, 2011) mentioned above explicitly reasons about support and occlusion relations between objects. But a more generic solution of integrating the semantics (together with uncertainties) of individual processing results still seems far off.

Dealing with Failure. A third, somewhat related reason is that researchers in individual specialised sub-fields (quite naturally) strive for perfection, inching recognition rates on standardised benchmarks ever higher in increments of 0.5%, while from a systems perspective it makes more sense to accept the inevitable uncertainties and failure modes and reason explicitly about them. This of coures hinges on having a common framework as explained above, to meaningfully express these unertainties. Even more importantly however it requires researchers to accept that perfection is futile.

4. What Vision Should Do

So what should be done to alleviate the above problems? There is of course no simple answer to that. But let us first look at some of the apparent solutions.

It isn't 3D. With the availability of cheap and powerful 3D sensors such as complete stereo solutions by companies like Point Grey [3] or Videre Design [4] or depth sensors such as the Mesa Imaging Swissranger [5] or Microsoft Kinect [6] one important part of vision seems to be have been solved, namely reconstructing a 3D scene. There is more to it though, as humans do not perceive a scene as a sort of flip-up cardboard diorama

[3] http://www.ptgrey.com/products/stereo.asp
[4] http://www.videredesign.com
[5] http://www.mesa-imaging.ch
[6] http://www.xbox.com/en-US/kinect

with missing object back sides. Reasoning about occluded parts of the scene as well as segmentation into individual objects remains to be solved.

More importantly however, a quick test on yourself by closing one eye will reveal that 3D sensing is not all that important for human vision. Picking up a small object or putting a key into a lock might require several attempts, so clearly direct perception of distance via stereo is an advantage for close-range manipulation, such as using tools or grasping branches when swinging from tree to tree. But the vivid impression of being situated in a 3D scene does not suffer significantly when being deprived of stereo vision. Also many grazing animals tend to have non-overlapping fields of view of the left and right eye, as large field of view (to notice approaching predators) is more important than accurate 3D perception.

For various reasons cups and cows are prominent example objects in computer vision. Advocates of 3D computer vision will point out that given a cup with a picture of a cow printed on it, a 2D recogniser would be likely to rather recognise a (nicely textured) cow than a (probably untextured) porcelain cup, whereas a 3D shape based recogniser would correctly identify the cup. However, given a 2D image of a cup with a cow on it humans have no problem recognizing both, the cup and the fact that there is a picture of a cow printed on it.

We are not arguing that 3D sensing would not be a powerful cue, and in fact robotics is likely to benefit a lot from depth sensors in the near future. But 3D sensing does not seem to be essential for perceiving a 3D scene. Human vision has developed powerful computational mechanisms to infer a complete 3D scene from rather limited information. And these mechanisms are more important then a specific sensing modality.

It isn't Resolution. In a similar vein image resolution does not seem to be critical. Certainly nature has evolved foveated vision for a good reason. The combination of attentional mechanisms based on low resolution cues with saccades to salient image regions to be processed at high resolution is an important mechanism to optimise visual processing and keep the amount of information tractable. Likewise any computer vision task benefits significantly from the object of interest being shown large and centred in the image, rather than occupying a small image region somewhere in the scene. However, humans looking at a low resolution, say 640×480, image of a scene typically have no problem interpreting it correctly (otherwise watching TV would be rather confusing).

Moreover, experiments with rapid serial visual presentation (Thorpe and Imbert, 1989) have shown that humans are remarkably good at identifying objects and scenes at presentation times well below 200 *ms*, which leaves no time to perform any saccades. For example in (Intraub, 1981) subjects were able to identify pictures of a category ("look for a butterfly"), superordinate category ("look for an animal") or negative category ("look for a picture that is *not* of house furnishings and decorations") presented for only 114 *ms*.

So the human visual system can perform (at least a significant deal of) its processing within an instant without requiring to scan the image with the high resolution fovea.

It isn't (just) Bayes. There is no doubt that much progress in vision is owed to the adoption of probabilistic frameworks over crisp symbolic methods, which are typically too

brittle when confronted with the cluttered, uncertain, ambiguous real world. However sometimes the actual probabilities at the end of some lenghty mathematical argument are rather ad-hoc, say the number of matching edgels divided by the total number of edgels, or assumptions about uniform priors. The respective approaches still work fine, thanks to the exraordinary robustness of statistical methods. But the results from different processing modules, although supposedly derived within the same mathematical framwork, become difficult to compare to each other within a common system. Just using probablities is not enough. Care has to be taken, that they refer in the same way to the same underlying causes.

A different way to treat uncertainties, rather than aiming for precise 3D estimates plus a measure or remaining uncertrainty, could be to not aim for exactness in the first place. Instead one could use more qualitiative measures, such as surface A is behind surface B, which is an observation that can be established with high certainty over a wide range of actual distances. This sort of information might be sufficiently accurate for many types of actions, such as reaching for A. However, it is not clear whether the mathematics for this kind of reasoning over a complex 3D scene would actually turn out to be simpler than more traditional probability theory.

Back to the Roots?. Armed with the lessons learned along the way (and with considerably increased computing power), it might be worth reconsidering some of the early approaches to computer vision. These in general aimed at reconstructing a 3D scene from very impoverished visual information, such as edge images only. While simply taking a depth sensor would certainly provide a more direct route, attacking the harder problem, with all the modern mathematical machinery, is still worthwile. Partly because advances there are more likely to shed light on the computational principles underlying human vision. But also because, as pointed out above, the problem of dealing with incomplete and ambiguous information persists, no matter how rich the underlying sensory information. This problem can be pushed back a little by more advanced sensors, but not avoided alltogether.

A Conjecture: Vision as Prediction. In the following we will put forward a conjecture of what might be one of the computational principles underlying human vision, based on an anecdotal example of severly impoverished visual information.

> I enter a room at night, put a glass of water on a table, walk back to the door to switch off the lights and the room becomes almost completely dark. I walk back to the table and can not actually see the glass or anything else on the table, or even the table itself. Scene reconstruction in this case is simply hopeless. Still I expect the glass at the same position where I left it and I can very roughly estimate that position by backtracking my steps. So I turn my head this way and that looking towards a window, which is slightly illuminated from outside, until I can see a glint typical of glass surfaces near the expected position. I reach out (carefully, as I might still collide with other unseen objects on the table) and successfully grab the glass (relying of course heavily on tactile feedback). By no means could I have reconstructed the glass with its 3D shape in that case, still I did "see"

it. Or rather I saw something I expected to see given that the glass were there.

Vision as a process of reconstructing the 3D scene is an ill-posed problem, yet humans seem to do it effortlessly. Still there are enough every-day cases where also for humans scene reconstruction becomes quite impossible (e.g., in very low light situations). Humans can however still employ vision successfully in such cases, because what if not vision eventually allowed the detection of the glass in the above darkened room example. Only a little visual information was needed to confirm some hypothesis about the scene.

The point is that that while reconstruction is a notoriously difficult problem, the inverse problem - prediction - is often very simple. One avenue of progress might thus be to view vision (at least in part) as a prediction problem, based on strong priors. A general framework should be able to incorporate multiple cues (visual and possibly non-visual), where the appearance of each cue (such as edges, shadows, highlights) is predicted, given a scene hypothesis. Predictions and actual observations could then used in a Bayesian filter to update an estimate of the scene.

This is just a rough sketch of course. But maybe observing human performance in such visually challenging situations can point the way to technical solutions that degrade equally gracefully.

5. Conclusion

The power of biological vision systems seems not to lie in perfect sensors and processing results, but in dealing with imperfect ones. Failures, uncertainties and ambiguities are not exceptional states of an otherwise perfectly functioning system, but instead part of the normal flow of processing.

We should thus aim at understanding the powerful computational principles allowing biologial vision to infer a sufficiently accurate model of reality from partial, ambiguous, sometimes erroneous information derived from various cues. Actually this is already happening within many of the approaches presented above in the form of probabilistic models, however not on a system wide level.

Bringing the individual successful pieces of vision together into an equally successful system that eventually allows robots to operate within the challenging environments of our appartments, remains a ambitious goal.

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., Szeliski, R., 2009. Building Rome in a Day. In: Proceedings of the International Conference on Computer Vision. pp. 72–79.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU) 110 (3), 346–359.

Biederman, I., 1987. Recognition-by-components: A Theory of Human Image Understanding. Psychological Review 94 (2), 115–147.

Biegelbauer, G., Vincze, M., Wohlkinger, W., 2010. Model-based 3D Object Detection: Efficient Approach Using Superquadrics. Machine Vision and Applications 21, 497–516.

Binford, T. O., 1971. Visual Perception by Computer. In: Proceedingsof the IEEE Conference on Systems and Control.

Björkman, M., Kragic, D., 2010. Active 3D Scene Segmentation and Detection of Unknown Objects. In: 2010 IEEE International Conference on Robotics and Automation (ICRA). pp. 3114–3120.

Brooks, R., 1983. Model-Based 3-D Interpretations of 2-D Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 5 (2), 140–150.

Chestnutt, J., Kagami, S., Nishiwaki, K., Kuffner, J., Kanade, T., 2007. GPU-Accelerated Real-Time 3D Tracking for Humanoid Locomotion. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.

Choi, C., Christensen, H., 2010. Real-Time 3D Model-Based Tracking Using Edge and Keypoint Features for Robotic Manipulation. In: IEEE Int. Conf. on Robotics and Automation. pp. 4048–4055.

Clowes, M. B., 1971. On seeing things. Artificial Intelligence 2 (1), 79–116.

Collet, A., Berenson, D., Srinivasa, S. S., Ferguson, D., 2009. Object Recognition and Full Pose Registration from a Single Image for Robotic Manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 3534–3541.

Cummins, M., Newman, P., November 2010. Appearance-only SLAM at large scale with FAB-MAP 2.0. The International Journal of Robotics Research.

Dalal, N., Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. In: IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2. pp. 886–893, introduces the HOG descriptor.

Davison, A., Reid, I., Molton, N., Stasse, O., 2007. Monoslam: Real-Time Single Camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6), 1052–1067.

Dickinson, S., 2009. The Evolution of Object Categorization and the Challenge of Image Abstraction. In: Dickinson, S., Leonardis, A., Schiele, B., Tarr, M. (Eds.), Object Categorization: Computer and Human Vision Perspectives. Cambridge University Press, pp. 1–37.

Dickinson, S., Pentland, A., Rosenfeld. A., 1992. 3-D Shape Recovery Using Distributed Aspect Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2), 174–198.

E. Rosten, T. Drummond, 2006. Machine Learning for High-Speed Corner Detection. In: Prococeedings of the 9th European Conference on Computer Vision (ECCV). pp. 430–434, introduces the FAST corner detector.

Fei-Fei, L., Fergus, R., Perona. P., 2006. One-shot Learning of Object Categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4), 594–611.

Ferrari, V., Tuytelaars, T., Van Gool. L. J., 2006. Simultaneous object recognition and segmentation from single or multiple model views. International Journal of Computer Vision 67 (2), 159–188.

Fitzgibbon, A. W., Zisserman, A., 2000. Multibody Structure and Motion: 3-D Reconstruction of Independently Moving Objects. In: Proceedings of the European Conference on Computer Vision. Springer-Verlag, pp. 891–906.

Fitzpatrick, P., Metta, G., 2003. Grounding vision through experimental manipulation. Philosophical Transactions: Mathematical, Physical and Engineering Sciences 361 (1811), 2165–2185.

Gordon, I., Lowe, D. G., 2006. What and where: 3D object recognition with accurate pose. In: Ponce, J., Hebert, M., Schmid. C., Zisserman, A. (Eds.), Toward Category-Level Object Recognition. Springer, Ch. What and where: 3D object recognition with accurate pose, pp. 67–82.

Hager, G. D., Wegbreit, B., 2011. Scene parsing using a prior world model. The International Journal of Robotics Research.

Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V., 2011. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: IEEE International Conference on Computer Vision (ICCV).

Hoiem, D., Efros, A., Hebert, M., 2006. Putting Objects in Perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2137–2144.

Huffman, D., 1971. Impossible Objects as Nonsense Sentences. Machine Intelligence 6.

Intraub, H., 1981. Rapid conceptual identification of sequentially presented pictures. Journal of Experimental Psychology: Human Perception and Performance 7, 604–610.

Klein, G., Murray, D., Sep. 2006. Full-3D Edge Tacking with a Particle Filter. In: Proc. British Machine Vision Conference (BMVC). Vol. 3. pp. 1119–1128.

Klein, G., Murray, D., 2007. Parallel tracking and mapping for small AR workspaces. In: Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR). Nara, Japan, pp. 225–234.

Kraft, D., Pugeault, N., Baseski, E., Popovic, M., Kragic, D., Kalkan, S., Wörgötter, F., Krüger, N., 2008. Birth of the object: Detection of objectness and extraction of object shape through object action complexes. International Journal of Humanoid Robotics 5 (2), 247–265.

Leibe, B., Schiele, B., 2003. Interleaved Object Categorization and Segmentation. In: Proceedings of the British Machine Vision Conference.

Lepetit, V., Fua, P., 2005. Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. Foundationas and Trends in Computer Graphics and Vision 1 (1), 1–89.

Lourakis, M. A., Argyros, A., 2009. SBA: A Software Package for Generic Sparse Bundle Adjustment. ACM Trans. Math. Software 36 (1), 1–30.

Lowe, D. G., 1987. Three-dimensional object recognition from single two-dimensional images. Artificial Intelligence 31 (3), 355–395.

Lowe, D. G., September 1999. Object Recognition from Local Scale-Invariant Features. In: International Conference on Computer Vision. pp. 1150–1157.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60 (2), 91–110.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W. H. Freeman.

Marr, D., Nishihara, H., 1978. Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. Proceedings of the Royal Society of London B 200 (1140), 269–294.

Matas, J., Chum, O., Martin, U., Pajdla, T., 2002. Robust wide baseline stereo from maximally stable extremal regions. In: Proceedings of the British Machine Vision Conference. Vol. 1. pp. 384–393.

Mikolajczyk, K., Schmid, C., 2004. Scale & Affine Invariant Interest Point Detectors. International Journal of Computer Vision 60 (1), 63–86.

Mörwald, T., Kopicki, M., Stolkin, R., Wyatt, J., Zurek, S., Zillich, M., Vincze, M., 2011. Predicting the Unobservable: Visual 3D Tracking with a Probabilistic Motion Model. In: Proceeedings of the IEEE International Conference on Robotics and Automation. pp. 1849–1855.

Mörwald, T., Prankl, J., Richtsfeld, A., Zillich, M., Vincze, M., 2010. BLORT - The Blocks World Robotic Vision Toolbox. In: Best Practice in 3D Perception and Modeling for Mobile Manipulation (in conjunction with ICRA 2010).

Murphy-Chutorian, E., Trivedi, M. M., 2008. Particle Filtering with Rendered Models: A Two Pass Approach to Multi-Object 3D Tracking with the GPU. In: CVPR workshop on Computer Vision on GPU's (CVGPU). pp. 1–8.

Nevatia, R., Binford, T. O., 1977. Description and Recognition of Curved Objects. Artificial Intelligence 8, 77–98.

Newcombe, R., Davison, A., 2010. Live Dense Reconstruction with a Single Moving Camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1498–1505.

Nistér, D., Naroditsky, O., Bergen, J., 2006. Visual Odometry for Ground Vehicle Applications. Journal of Field Robotics 23 (1).

Ozden, K. E., Schindler, K., Gool, L. V., 2010. Multibody structure-from-motion in practice. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1134–1141.

Özuysal, M., Fua, P., Lepetit, V., 2007. Fast Keypoint Recognition in Ten Lines of Code. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8.

Özuysal, M., Lepetit, V., Fleuret, F., Fua, P., 2006. Feature Harvesting for Tracking-by-Detection. In: European Conference on Computer Vision. Vol. 3953. pp. 592–605.

Pan, Q., Reitmayr, G., Drummond, T., 2009. ProFORMA: Probabilistic Feature-based On-line Rapid Model Acquisition. In: Proc. British Machine Vision Conference (BMVC). pp. 1–11.

Pilet, J., Lepetit, V., Fua, P., January 2007. Fast non-rigid surface detection, registration and realistic augmentation. International Journal of Computer Vision.

Roberts, L. G., 1965. Machine perception of three-dimensional solids. In: Tippett, J. T. (Ed.), Optical and Electro-Optical Information Processing. MIT Press, Cambridge, MA, pp. 159–197.

Rusu, R. B., Blodow, N., Marton, Z. C., Beetz, M., 2009. Close-Range Scene Segmentation and Reconstruction of 3D Point Cloud Maps for Mobile Manipulation in Human Environments. In: Proceedsings of the 22nd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1–6.

Sánchez, J., Álvarez, H., Borro, D., 2010. Towards Real Time 3D Tracking and Reconstruction on a GPU using Monte Carlo Simulations. In: 9th IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 185–192.

Snavely, N., Seitz, S. M., Szeliski, R., 2006. Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings. pp. 835–846.

Thorpe, S. J., Imbert, M., 1989. Biological constraints on connectionist modelling. In: Connectionism in Perspective. Elsevier, pp. 63–92.

Thrun, S., Burgard, W., Fox, D., 2005. Probabilistic Robotics. MIT Press.

Tola, E., Lepetit, V., Fua, P., 2008. A Fast Local Descriptor for Dense Matching. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8.

Waltz, D., 1975. Understanding Line Drawings of Scenes with Shadows. In: Winston, P. H. (Ed.), The Psychology of Computer Vision. McGraw-Hill, New York, pp. 19–91.

Welke, K., Issac, J., Schiebener, D., Asfour, T., Dillmann, R., 2010. Autonomous Acquisition of Visual Multi-View Object Representations for Object Recognition on a Humanoid Robot. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 2012–2019.

Zach, C., Pock, T., Bischof, H., 2007. A Duality Based Approach for Realtime TV-L1 Optical Flow. Pattern Recognition 4713, 214–223.