Chapter 1
Recent Trends in Computational and Robot Vision

Ville Kyrki and Danica Kragic

1.1 Introduction

There are many characteristics in common in computer vision research and vision research in robotics. For example, the Structure-and-Motion problem in vision has its analog of SLAM (Simultaneous Localization and Mapping) in robotics, visual SLAM being one of the current hot topics. Tracking is another area seeing great interest in both communities, in its many variations, such as 2-D and 3-D tracking, single and multi-object tracking, rigid and deformable object tracking. Other topics of interest for both communities are object and action recognition.

Despite having these common interests, however, ”pure” computer vision has seen significant theoretical and methodological advances during the last decade which many of the robotics researchers are not fully aware of. On the other hand, the manipulation and control capabilities of robots as well as the range of application areas have developed greatly. In robotics, vision can not be considered an isolated component, but it is instead a part of a system resulting in an action. Thus, in robotics the vision research should include consideration of the control of the system, in other words, the entire perception-action loop. A holistic system approach would then be useful and could provide significant advances in this application domain.

We believe that although there have been good examples of robust vision systems, there is a gap between the research conducted in computer vision and robotics communities. In the following, we aim to identify some of the recent developments where we see great potential for the co-operation between the communities.

Ville Kyrki
Lappeenranta University of Technology, Department of Information Technology, P.O. Box 20, 53851 Lappeenranta, Finland, e-mail: kyrki@lut.fi

Danica Kragic
Centre for Autonomous Systems, Computational Vision and Active Perception, School of Computer Science and Communication, Royal Institute if Technology, 10044 Stockholm, Sweden, e-mail: dani@kth.se

1.2 Perception and Action

One of the main challenges in the field of robotics is to make robots ubiquitous. To intelligently interact with the world, robots need to perceive and interpret the environment and situations around them and react appropriately. In other words, they need context-awareness. But how to equip robots with capabilities of gathering and interpreting the necessary information for novel tasks through interaction with the environment and by providing some minimal knowledge in advance? This has been a longterm question and one of the main drives in the field of cognitive system development. For a service robot that is to perform tasks in a human environment, it has to be able to learn about objects and object categories. However, the robots will not be able to form useful categories or object representations by being only passive observers of the environment. They should, like humans, learn about objects and their representations through interaction.

The simplest type of interactions that can occur between a robot and an object may be to, for example, push an object in order to retrieve information about the size or weight of the object. Here, simple visual cues providing approximate 3D position of the object may be sufficient. A more complex interaction may be to grasp the object for the purpose of gaining the physical control over the object. Once the robot has the object in its hand, it can perform further actions on it, such as examining it from other views. Information obtained during interaction can be used to update the robots representations about objects and the world.

Such an approach is studied in Chapter 2 where a mobile manipulator system employs interactive perception to extract kinematic models from tools such as pliers or shears. The extracted models are then used to compute an action that transforms the kinematic attributes of the object to actions that can be performed on them thus mimicking tool use.

Another way of learning of how to interact with the environment is to observe humans or other agents and perform actions through a process of *imitation*, [4]. Imitation learning in robotics is therefore strongly related to action representation and recognition. The overall goal of imitation learning is to develop robots that are able to relate perceived actions of another agent to its own embodiment in order to learn, recognize and finally perform the demonstrated actions [8, 40, 41, 19, 29, 30, 22, 11, 6]. Most of the work work on imitation in robotics is motivated by human findings, where there is strong neurobiological evidence that human actions and activities are directly connected to the motor control of the human body [14, 36, 37].

Detection of human body and body part, recognition and interpretation of their actions has therefore during the past few years gained considerable amount of attention, [1, 2, 13, 48, 24, 47]. Main motivation is the large number of potential applications, e.g., in visual surveillance, entertainment industry, robot learning and control due to the ability to acquire and store a large amount of data that can be processed offline.

In visual surveillance applications, the work is mainly concentrated on classification of common versus uncommon actions. In the entertainment industry, the interest lies mainly in the field of motion capture and synthesis where precise motion capture allows to replace an actor with a digital avatar. Ideally, the motion capture system should be non-intrusive thus making vision based techniques a natural solution.

Even if there are differences in the applications, the system design is very similar: the sensory input has to be acquired and represented so to enable i) the recognition of the observed actions and ii) understand the effects certain actions may have on the environment. In robotics, there is an additional dimension since the system also has to enable the robot to physically perform a certain action in order to cause the desired change in the environment. The last point point depends on the individual/robot under consideration: how to perform an action that causes a particular change in the environment for a human and a robot depends on their physical capabilities.

Chapter 3 presents an approach for people detection with a mobile robot through combination of visual and laser range scanner sensors. It is shown how a person may be represented as a constelation of body parts to facilitate the detection process.

Chapter 4 presents a system where the road towards developing cognitive capabilities in robots is followed by considering an interply between perception and action. A humanoid robot with perception, manipulation and communication skills is described with a special attention paid on the design of the robot head as the base for prividing the input to various visuo-motor behaviors. The presented work spans from human motion tracking and object representation to action imitation and adaptation.

One of the important aspects of integrating the results from the computer vision community on mobile robots that have constrained processing and storage capabilities, is to consider how to adapt complex processing methods for real-time applications. In Chapter 5, it is shown how time-constrained classification, detection and matching problems may be formalised in the framework of sequential decision-making. The work derives quasi-optimal time-constrained solutions for the three problems of major relevance for the community: feature detection, feature mathcing and face detection.

Chapter 6 presents an application where the real-time processing aspects are of utmost importance - a medical application intended for 3D positioning and guidance of surgical instruments in the human body. The work shows the importance of real-time visual processing for the purpose of vision guided control where model based techniques are adopted to facilitate the tracking process.

1.3 Mapping the Environment — SLAM and vSLAM

One of the essential capabilities of an autonomous mobile robot is to move around in its environment. To accomplish this in complex natural environments, the robot needs the ability to build maps of the environment using natural landmarks and to use them for localization [44, 7, 10, 43, 45]. One of the current research topics related to Simulatanous Localization and Mapping (SLAM) is the use of vision as the only exteroceptive sensor [9, 12, 15, 42, 27], due to its low cost. We adopt the term vSLAM[17] for visual SLAM. Currently, vSLAM solutions focus on accurate localization, mostly based on the estimation of geometric visual features of the environment. Thus, the resultant map is useful for the localization of the robot, but its use for other purposes is often neglected. This section concentrates on the geometric mapping while in the next section, we try to take a look at some opportunities how to apply visual means for higher-level understanding of the robot's environment.

In mainstream computer vision research, the problem of 3D reconstruction from a sequence of images is termed structure from motion. There is a great similarity between the structure from motion (SfM)[1] and vSLAM. The essential problem of simultaneously estimating the structure of the environment and the motion of the observer is identical. In computer vision community, SfM is nowadays considered mostly a solved problem, as commercial solutions for SfM-based camera motion estimation have become available from companies such as 2d3[2] The state-of-art SfM solutions are mostly based on using projective geometry as the geometrical model [16] and bundle adjustment techniques (basically Levenberg-Marquardt minimization) for finding the maximum likelihood solution for the nonlinear optimization problem [16, 46]. In addition to the 3D reconstruction for calibrated cameras, the internal camera calibration can be estimated along with the structure (self calibration), thus eliminating the need to explicitly calibrate cameras (for example, [33, 21]).

However, there are some differences between SfM and vSLAM which are often overlooked. The main emphasis in SfM is to reconstruct the structure of the environment as accurately as possible. This is often termed "global bundle adjustment", emphasizing the fact that the maximum likelihood solution over the whole image sequence is sought. Thus, the solutions are essentially batch algorithms, requiring the whole data for processing. Also, the heavy computational load of the global optimization makes it impossible to run in real-time. One reason for this is that the intended application areas, such as camera tracking for the movie industry, do not require on-line estimation. In vSLAM, on-line operation is a fundamental requirement, as the information is often used for robot navigation. In practice, the computational complexity increases with respect to the number of landmarks and time. If the robot needs to operate in a large (or changing) environment for a long time, the vSLAM algorithm would need to be constant time (or at least sub-linear). Currently there is no vSLAM approach which would allow this, although there have been propositions of constant time SLAM algorithms using other sensors (for example, [20]) although the estimates given by the methods are conservative rather than optimal.

Another characteristic typical for vSLAM but not for SfM is that in addition to the estimate, its uncertainty needs to be characterized. Moreover, the algorithms need to be statistically consistent, that is, not giving overly confident estimates of the uncertainty. This characteristic can be exploited to build safeguards and allow the information to be safely used for controlling the robot.

Loop closing refers to identifying that a robot has returned to a previously visited area after touring in another one. The observation of previously known landmarks

[1] Nowadays often termed "structure and motion."

[2] See `http://www.2d3.com`.

allows to improve the estimate over the whole tour. Loop closing using the traditional SLAM sensors, for example, laser scanners, is hard due to the difficulties in correctly matching previously detected landmarks and current observations. Visual features are very powerful in this respect, and the use of visually salient features for loop closing has been recently proposed [26, 34]. However, there are still many unexplored possibilities in using image retrieval and matching approaches for loop closing.

In some cases, the maps built are not intended (or needed) to be used for localization at later time instants. In this case, the resulting map is only stored for a short period and the landmarks are removed from the map after they are no longer visible. This approach is usually called visual odometry [28, 3]. While this approach essentially solves the map complexity problem, it suffers from drift as the earlier landmarks are forgotten and loop closing is not possible. As a compromise between the accuracy of global bundle adjustment and the restrictions in the available processing capacity, it has been proposed to only optimize the pose of a small number of previous time instants [25]. This has the effect that the accuracy is increased, but unfortunately the local approach still can not remove drift and benefit from loop closing in contrast to the more global approaches.

Ideally, the robotics community would like to get a statistically consistent vSLAM algorithm that would perform constant time on-line estimation with the optimality of the global bundle adjustment, allowing efficient loop closing. This book contains three chapters which demonstrate some of the ideas and recent work towards this ideal goal. Chapter 7 presents the Sliding Window Filter, an approach for incremental SLAM using a sliding time window of most recent sensor measurements, thus attaining constant time complexity. Chapter 8 examines the joint use of object/scene recognition for coarse topological localization and more accurate local metric localization using 1D trifocal tensor. Chapter 9 discusses different types of visual landmarks which could be used for vSLAM. Furthermore, loop closing and future directions such as multi-robot SLAM and the use of geographical information systems (GIS) in SLAM are considered.

1.4 From Maps to Understanding the Environment

Robots of the future should be able to easily navigate in dynamic and crowded environments, detect as well as avoid obstacles, have a dialog with a user and manipulate objects. It has been widely recognized that, for such a system, different processes have to work in synergy: high-level cognitive processes for abstract reasoning and planning, low-level sensory-motor processes for data extraction and action execution, and mid-level processes mediating these two levels.

A successful coordination between these levels requires a well defined representation that facilitates anchoring of different processes. One of the proposed modeling approaches has been the use of *cognitive maps* [5]. The cognitive map is the body of knowledge a human or a robot has about the environment. In [5], it is argued that topological, semantic and geometrical aspects are important for representation of spatial knowledge. This approach is closely related to Human-Augmented mapping (HAM) where a human and a robot interact so to establish a correspondence between the human spatial representation of the environment and robot's autonomously learned one [18].

In addition, both during the mapping phase and during robot task execution, object detection can be used to augment the map of the environment with objects' locations, [39]. There are several scenarios here: while the robot is building the map it will add information to the map about the location of objects. Later, the robot will be able to assist the user when s/he wants to know where a certain object X is. As object detection might be time consuming, another scenario is that the robot builds a map of the environment first and then when no tasks are scheduled for execution, it moves around in the environment and searches for objects.

The same skill can also be used when the user instructs the robot to go to a certain area to fetch a particular object. If the robot has seen the object before and has it already in the map, the searching process is simplified to re-detection. By augmenting the map with the location of objects we also foresee a way of achieving place recognition. This provides valuable information to the localization system as well as it greatly reduces the problem with symmetries in a simple geometric map. This would be an alternative approach to the visual place recognition presented in [35] and the laser based system in [23]. Furthermore, along the way by building up statistics about what type of objects typically can be found in, for example, a kitchen the robot might not only be able to recognize a certain kitchen but also potentially generalize to recognize a room it has never seen before as probably being a kitchen.

However, although there exists a large body of work on mobile robots, there are still no fully operational systems that can operate robustly and long-term in everyday environments. The current trend in development of service robots is reductionistic in the sense that the overall problem is commonly divided into manageable subproblems.

Chapter 10 explores the relations between tasks, objects and contexts for robots using maps, in the context of visually guided robots. Task descriptions are necessary as they explain the structure of actions and objects the tasks require. Most tasks can only occur in prototypical places, which have a suitable arrangement of objects. The local set of objects and their configuration then determines the context. The context then allows to endow physical locations with semantic meaning.

A solution to the visual SLAM problem does not necessarily allow robot navigation. Specifically, the knowledge of the 3-D location of a robot is insufficient for solving the navigation in a general setting. On the other hand, navigation can also be possible using solely image-based information. Chapter 11 investigates the use of visual information for robot navigation, presenting both image-based and map-based approaches.

1.5 Future Directions

The recent trends and future directions in the computer vision community are not always well-known for robotics researchers. This section tries to outline some of them[3].

Benchmarking is an important issue, which has only lately gained notable interest in robotics[4]. Since early 1990s there has been a growing interest in the computer vision community on discovering ways to compare the performance of different methods. For example, the EU funded PCCV (Performance Characterization in Computer Vision) project produced tutorials and case-studies for benchmarking vision algorithms [31]. Based on these studies, important benchmarking techniques include common benchmark data, contests, and specialized benchmarking workshops. While all of these have begun to appear also in robotics [32], it seems that the robotics community could benefit from the issues learned. Another recent trend in validating methods in computer vision is to use huge data sets gathered from internet. For example, image search engines give a possibility to obtain thousands of labeled images for testing object and scene recognition approaches.

Machine learning methods are becoming more and more widely used in computer vision as the processing power of modern computers has reached the point to make this possible. Most importantly, they are not only used for traditional "recognition" applications, but they can be also used for constructing efficient and effective processing algorithms, for example, for image feature extraction. Chapter 5 of this book is a fine example of this development, as is Rosten's work in learning high-speed corner detection [38].

Robotics has for the past few decades evolved from an industrial, well-controlled environment, to our homes, medical/operating rooms and resulted in sending robots to Mars. Still, most of the existing robot systems are designed for specific purposes and preprogrammed to expected requirements posed by the task and the environment. As mentioned, the challenge for the future is to go beyond the current engineering paradigm and develop artificial cognitive robotic systems that can robustly perceive, interact, reason and cooperate with humans and each other in open-ended environments.

Compared to classical robot systems, where the main requirement of the system is to execute a predefined task, the problem of task learning and planning stands as an open problem. Easy and user-friendly programming of new tasks in robots is one of the integral problems that will have to be considered in the future service robot systems.

The work presented in Chapter 12 takes a step towards solving this problem by providing a programming support for the implementation of complex, sensor-based robotic tasks in the presence of geometric uncertainty. The application of the proposed framework is studied in image-based visual servoing tasks.

[3] Some of the following ideas originate from a panel discussion at the ICRA 2007 workshop, which was a source of inspiration for this book.

[4] A benchmarking initiative has been recently started in EURON, the European robotics network.

Acknowledgements The support to Ville Kyrki by Academy of Finland is gratefully acknowledged.

References

1. Aggarwal, J., Park, S.: Human Motion: Modeling and Recognition of Actions and Interactions. In: Second International Symposium on 3D Data Processing, Visualization and Transmission. Thessaloniki, Greece (2004)
2. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding: CVIU **73**(3), 428–440 (1999)
3. Agrawal, M., Konolige, K.: Rough terrain visual odometry. In: International Conference on Advanced Robotics. Jeju, Korea (2007)
4. Billard, A.: Imitation: A review. Handbook of brain thory and neural network, M. Arbib (ed.) pp. 566–569 (2002)
5. B.J.Kuipers: The cognitive map: Could it have been any other way? In H. L. Pick, Jr. and L. P. Acredolo (Eds.), Spatial Orientation: Theory, Research, and Application, New York: Plenum Press, pp. 345–359 (1983)
6. Calinon, S., Guenter, F., Billard, A.: Goal-Directed Imitation in a Humanoid Robot. In: International Conference on Robotics and Automation. Barcelona, Spain, April 18-22 (2005)
7. Castellanos, J.A., Tardós, J.D.: Mobile Robot Localization and Map Building: A Multisensor Fusion Approach. Kluwer Academic Publishers (1999)
8. Dariush, B.: Human Motion Analysis for Biomechanics and Biomedicine. Machine Vision and Applications **14**, 202–205 (2003)
9. Davison, A.: Real-time simultaneous localisation and mapping with a single camera. In: Internatinal Conference on Computer Vision (2003)
10. Dissanayake, G., Newman, P., Clark, S., Durrant-Whyte, H., Corba, M.: A solution to the slam building problem. IEEE Transactions on Robotics **17**(3), 229–241 (2001)
11. Ekvall, S., Kragic, D.: Grasp recognition for programming by demonstration tasks. In: IEEE International Conference on Robotics and Automation, ICRA'05, pp. 748 – 753 (2005)
12. Folkesson, J., Jensfelt, P., Christensen, H.: Vision slam in the measurement subspace. In: International Conference on Robotics and Automation (2005)
13. Gavrila, D.M.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding: CVIU **73**(1), 82–98 (1999)
14. Giese, M., Poggio, T.: Neural Mechanisms for the Recognition of Biological Movements. Nature Reviews **4**, 179–192 (2003)
15. Goncavles, L., di Bernardo, E., Benson, D., Svedman, M., Ostrovski, J., Karlsson, N., Pirjanian, P.: A visual front-end for simultaneous localization and mapping. In: International Conference on Robotics and Automation, pp. 44–49 (2005)
16. Hartley, R., Zisserman, A.: Multiple View Geometry. Cambridge University Press (2003)
17. Karlsson, N., di Bernardo, E., Ostrowski, J., Goncalves, L., Pirjanian, P., Munich, M.: The vSLAM algorithm for robust localization and mapping. In: International Conference on Robotics and Automation, pp. 24–29. Barcelona, Spain, April 18-22 (2005)
18. Kruijff, G.J.M., Zender, H., Jensfelt, P., Christensen, H.I.: Clarification dialogues in human-augmented mapping. In: Proc. of the 1st Annual Conference on Human-Robot Interaction, HRI'06. Salt Lake City, UT (2006)
19. Kuniyoshi, Y., Inaba, M., Inoue, H.: Learning by watching, extracting reusable task knowledge from visual observation of human performance. In: IEEE Transactions on Robotics and Automation, vol. 10(6), pp. 799–822 (1994)
20. Leonard, J., Newman, P.: Consistent, convergent, and constant-time SLAM. In: International Joint Conference on Artificial Intelligence, pp. 1143–1150. Acapulco, Mexico (2003)

21. Lhuillier, M., Quan, L.: A quasi-dense approach to surface reconstruction from uncalibrated images. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(3), 418–433 (2005)
22. Lopes, M.C., Victor, J.S.: Visual transformations in gesture imitation: What you see is what you do. In: International Conference on Robotics and Automation, pp. 2375– 2381 (2003)
23. Martínez Mozos, O., Stachniss, C., Burgard, W.: Supervised learning of places from range data using adaboost. In: Proc. of the IEEE International Conference on Robotics and Automation, ICRA'05, pp. 1742–1747. Barcelona, Spain (2005)
24. Moeslund, T., Hilton, A., Krueger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding **104**(2-3), 90–127 (2006)
25. Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P.: Real-time localization and 3D reconstruction. In: Computer Vision and Pattern Recognition. New York City, New York, USA, June 17-22 (2006)
26. Newman, P., Ho, K.: Slam- loop closing with visually salient features. In: International Conference on Robotics and Automation. Barcelona, Spain, April 18-22 (2005)
27. Newman, P., Ho, K.: SLAM-loop closing with visually salient features. In: International Conference on Robotics and Automation, pp. 644–651 (2005)
28. Nister, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Computer Vision and Pattern Recognition, pp. 652–659. Washington DC, USA, June (2004)
29. Ogawara, K., Iba, S., Kimura, H., Ikeuchi, K.: Recognition of human task by attention point analysis. In: International Conference on Intelligent Robots and Systems, pp. 2121–2126 (2000)
30. Ogawara, K., Iba, S., Kimura, H., Ikeuchi, K.: Acquiring hand-action models by attention point analysis. In: International Conference on Robotics and Automation, pp. 465–470 (2001)
31. PCCV. Performance Characterization in Computer Vision website. Http://peipa.essex.ac.uk/benchmark/index.html
32. del Pobil, A.P. (ed.): Benchmarks in Robotics Research. IROS 2006 workshop (2006)
33. Pollefeys, M., Koch, R., Van Gool, L.: Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. International Journal of Computer Vision **32**(1), 7–25 (1999)
34. Posner, I., Schroeter, D., Newman, P.: Using scene similarity for place labeling. In: International Symposium on Experimental Robotics. Rio de Janeiro, Brazil (2006)
35. Pronobis, A., Caputo, B., Jensfelt, P., Christensen, H.: A discriminative approach to robust visual place recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'06 (2006)
36. Rizzolatti, G., Fogassi, L., Gallese, V.: Parietal Cortex: from Sight to Action. Current Opinion in Neurobiology **7**, 562–567 (1997)
37. Rizzolatti, G., Fogassi, L., Gallese, V.: Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action. Nature Reviews **2**, 661–670 (2001)
38. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: European Conference on Computer Vision, vol. 1, pp. 430–443 (2006)
39. S. Ekvall, P.J., Kragic, D.: Object detection and mapping for service robot tasks. Robotics **25(2)**, 175–188, (2007)
40. Schaal, S.: Is Imitation Learning the Route to Humanoid Robots? Trends in Cognitive Sciences **3**(6), 233–242 (1999)
41. Schaal, S., Ijspeert, A., Billard, A.: Computational approaches to motor learning by imitation. Philosophical transaction of the Royal Society of London, series B **358**(1431), 537–547 (2003)
42. Sim, R., Elinas, P., Griffin, M., Little, J.J.: Vision-based slam using the rao-blackwellised particle filter. In: IJCAI Workshop on Reasoning with Uncertainty in Robotics (2005)
43. Tardós, J., Neira, J., Newman, P., Leonard, J.: Robust mapping and localization in indoor environments using sonar data. International Journal of Robotics Research **4** (2002)
44. Thrun, S., Fox., D., Burgard, W.: A probalistic approach to concurrent mapping and localization for mobile robots. Autonomous Robots **5**, 253–271 (1998)

45. Thrun, S., Liu, Y., D.Koller, Ng, A., Ghahramani, Z., Durrant-White, H.: SLAM with sparse extended information filters. International Journal of Robotics Research **23**(8), 690–717 (2004)
46. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment – A modern synthesis. In: W. Triggs, A. Zisserman, R. Szeliski (eds.) Vision Algorithms: Theory and Practice, LNCS, pp. 298–375. Springer Verlag (2000)
47. Veeraranghavan, A., Chellappa, R., Roy-Chowdhury, A.: The Function Space of an Activity. In: Computer Vision and Pattern Recognition. New York City, New York, USA, June 17-22 (2006)
48. Wu, Y., Huang, T.S.: Vision-based gesture recognition: A review. Lecture Notes in Computer Science **1739**, 103+ (1999)