Editorial

The International Journal of
Robotics Research
2015, Vol. 34(4-5) 399–401

DOI: 10.1177/0278364915574960
ijr.sagepub.com

Special Issue on Robot Vision

The International Journal of Robotics Research (IJRR) has a long history of publishing the state-of-the-art in the field of robotic vision. This is the fourth special issue devoted to the topic. Previous special issues were published in 2012 (Volume 31, No. 4), 2010 (Volume 29, Nos 2–3) and 2007 (Volume 26, No. 7, jointly with the *International Journal of Computer Vision*). In a closely related field was the special issue on *Visual Servoing* published in *IJRR*, 2003 (Volume 22, Nos 10–11). These issues nicely summarize the highlights and progress of the past 12 years of research devoted to the use of visual perception for robotics.

Looking back across these issues we see perennial topics such as calibration; feature detection, description and matching; multi-view geometry; and filtering and prediction. Of course for robotic vision we have also seen many papers with a strong control focus and also a focus on high-speed operation. Perennial challenges over that period, perhaps still open problems, include robustness and vision-guided manipulation. Happily, many techniques have matured over this period and become an integral part of many robotic vision systems, for example visual odometry, visual Simultaneous Localization and Mapping (SLAM), visual place recognition and the fusion of vision with other sensors, most notably inertial sensors.

This period has truly seen amazing technological change, not just the constant progress due to Moore's law but major innovations such as field-programmable gate arrays (FPGAs) and graphics processing units (GPUs), mobile computing architectures, low-cost high-performance inertial sensors and RGB-D sensors. Many of these have been driven by demand for consumer products such as smartphones and games, but have also provided a rich bounty for roboticists. The ready availability of capable low-cost off-the-shelf robotic platforms for domains such as underwater autonomous unmanned vehicles (AUVs), flying unmanned aerial vehicles (UAVs) and humanoid robots, all of which could usefully use vision sensors, is also helping to advance the field. Finally, the staple of all robotic vision systems, the camera, is evolving in very interesting directions. We now have cameras that are small, cheap and lightweight, that have progressive scan and global shutters, high dynamic range, high frame rate and wide fields of view obtained by catadioptrics or by multiple cameras with stitched imagery.

High-quality open-source software for robotic vision such as ROS and OpenCV has matured and gained wide currency over the period. The last few years has also seen machine learning techniques demonstrate ever-growing performance for a wide range of applications but particularly in vision-based problems such as object recognition and scene understanding.

Taken together these advances have significantly lowered the time and cost entry barriers for tackling and advancing important vision and robot perception problems. They have also opened avenues for new applications, enabled more sophisticated robotic tasks and more robust deployment of visual sensing in larger scale settings.

Interest in robotic vision is growing rapidly, as evidenced by the number of technical sessions in our leading conferences (ICRA, IROS, RSS and ISRR) and also by the number of submissions to this special issue. The call for papers for this issue was issued in late 2013 and a total of 47 papers were submitted (a record for the robot vision special issue). After a thorough review process we have accepted 15 papers for publication, which will appear as a double issue. Seven of these papers also feature multimedia contents.

The first two papers are concerned with humanoid robotics. Research in this area is now quite mature with respect to control issues and it is now time to close the perception–action loop. The first paper by Garcia, Stasse, Hayet, Dune, Esteves and Laumond proposes a visual servoing scheme to control dynamic walking of the HRP2 humanoid robot. The main idea of this paper is to incorporate the visual error (three-dimensional (3D) information in the proposed method) into a model-predictive control formulation for walking pattern generation. The approach proves to be more efficient than decoupling the computation of a reference velocity that is given by the pattern generator. The proposed approach is validated in simulation and we look forward to seeing the full validation of this control scheme on the actual HRP2 in future.

Calibration has always been an important issue in robot vision. The paper by Birbach, Frese and Bäuml addresses the problem of inertial measurement unit (IMU)–StereoCamera calibration with respect to the robot's kinematic chain (head, torso and wrist). As part of the calibration process, the intrinsic and extrinsic parameters of the sensors (stereo cameras, Kinect sensor, IMU) in the robot's head are determined. The calibration process is validated with the dynamic task of ball catching using onboard visual sensing as well as mobile manipulation.

For natural human–robot interaction (HRI) vision is important but other senses, sound and vision, warrant

consideration. Alameda-Pineda and Horaud propose combining vision-based HRI with robot hearing and audio-based HRI in an approach to detect and localize people that are both seen and heard. Localizing the speaker not only allows for better speech or speaker recognition by reducing the signal-to-noise ratio but also eases the tracking in the image stream. Audio-cues are generated by inter-aural time delay between stereo microphones, and visual signals are generated by stereo vision interest points and face detection. The system is extensively described and tested on publicly available datasets and the NAO humanoid robot. This work could be applied to the identification of human activity, for example, counting the number of speakers, localizing them, assessing their speaking state and so on.

As for previous special issues, many visual SLAM papers were submitted, and four are included here. SLAM is now routine for a vision-based robotics system and has now been successfully deployed into consumer robots. The SLAM papers published in this special issue not only provide methodological contributions but actual and efficient systems that are deployed in the field.

The paper by Kim and Eustice addresses a very important issue in SLAM and especially underwater SLAM, where visual features are sparse and imaging conditions challenging. It links perception and action in a meaningful way for localization and exploration problems. Whereas in a conventional approach, SLAM is passive and performed on preplanned trajectories, here the navigation is considered concurrently with the SLAM and is computed, using a decision theoretic approach, with two competitive objectives – efficient map coverage and minimum overlap – in order to avoid drift and bound the navigation error. A real-world experiment of this visual SLAM system for the task of autonomous underwater ship hull inspection is presented.

Camera localization is a fundamental issue in vision-based robotics. The paper by Lim, Sinha, Cohen, Uyttendaele and Kim details an algorithm for real-time localization of a moving monocular camera with respect to a previously generated 3D point cloud (using a structure from motion approach). A key originality of the approach is the ability to use the algorithm on an embedded computer with low computing power and to distribute computation time over several frames. Rather than considering online extraction of scale-invariant descriptors, an indexed database containing multiple descriptors per 3D point extracted at multiple scales is constructed offline. Computational efficiency is achieved by using these descriptors in an active matching approach during frame-to-frame tracking. The authors introduce several important new techniques to present a well thought out and very efficient system.

The work of Zhao, Huang, Sun, Yan and Dissanayake proposes a novel parameterization for bundle adjustment based on parallax angles for SLAM applications. The new modeling takes into account the parallax of the image point features rather than the more common depth parameter. They show that there are situations, such as when the camera motion is aligned with the direction of the features or when these features are far away, where existing parameterizations (such as the inverse depth parameterization) lead to poor conditioning in the optimization problem. The proposed formulation is shown to be much better conditioned, leading to improved motion and structure estimates. Results are presented for a number of standard datasets and open-source code is provided.

The paper by Kim, Yoon and Kweon also considers navigation using visual SLAM. With respect to the previous SLAM papers, the approach taken in this paper emphasizes key frame-based camera tracking and 3D mapping using a Bayesian filtering framework. The authors make the observation that, when using motion estimation from visual odometry to predict the new frame pose, the process noise is no longer independent of measurement noise if the same observations have been used in visual odometry and in SLAM. To solve this issue, that is, ensuring that process and measurement noises are independent, the available measurements are divided into two sets that are exclusively used by each process. The authors proceed by developing a Rao–Blackwellized particle filter formulation for this segmentation of the measurement set using, for computational efficiency, a FastSLAM-like process.

Robotic vision has definitively reached a mature level when a robot controlled by vision sensors can be deployed in the field. The paper by Hansen, Alismail, Rander and Browning deals with an important industrial issue related to gas pipeline inspection. A visual odometry-based system uses calibrated fish-eye imagery and sparse structured lighting to produce 3D textured surface models of the pipe's internal surface using a sparse bundle adjustment framework. This paper describes a well-integrated reconstruction system and significant experimentation is presented.

The work of Guizilini and Ramos proposes a technique for learning an automatic segmentation of static and dynamic objects in monocular sequences in an unsupervised setting. The approaches build upon techniques from monocular motion estimation and robust matching, and classify feature tracks as static or dynamic by incrementally training a Gaussian process classifier. This suggests a promising approach for learning models of semantically meaningful objects in an unsupervised setting, as well as enabling this strategy to attain reliable visual odometry estimates in the presence of moving objects.

The paper by Cadena and Košecká tackles the problem of simultaneous segmentation and categorization in RGB-D data, by effectively and efficiently segmenting the commonly occurring background classes in indoors environments such as ground, structure, furniture and props. This initial labeling is then sampled and predictions of the locations of generic object instances are made. The problem is cast in the Conditional Random Field framework and evaluated extensively in indoors environments. This paper presents a new avenue for obtaining hypotheses about previously unseen objects that can be exploited in object search and follow-up categorization.

The paper by Whelan, Kaess, Johannsson, Fallon, Leonard and McDonald describes a complete system to generate large-scale RGB-D sensor-based reconstructions of indoor environments. The proposed method is based on volumetric fusion of depth maps. The focus is on real time (30 Hz on a GPU) for all the component subtasks, such as camera tracking, mesh construction and pose graph optimization, which is achieved by making strong use of incremental updates. The paper presents impressive results for a state-of-the-art problem.

In the paper by Teo, Fermüller and Aloimonos a method for object detection and categorization using object shape in a cluttered indoor environment is proposed. The authors first detect likely location objects using the so-called torque operator, and this is followed by a novel representation of contours that enable effective multi-scale matching and grouping. The performance is evaluated on several diverse datasets containing a variety of object categories. The method is further improved when depth information is available from an RGB-D sensor. This approach is applicable to table-top setting categorization, can tolerate a large amount of clutter, occlusion and viewpoint changes and is effective for objects with discriminative shape characteristics.

The paper by Kim and Ueda deals with an important but little-studied problem of removing motion blur for fast-moving eyes. Saccades are very fast eye motions that allow humans to compensate for their limited field of view. Here a very fast lead zirconate titanate-driven pan-tilt camera system was employed to simulate such rapid eye motion. Motions are so fast that motion blur, caused by relative motion between the camera and an object, would prevent any possibility of scene interpretation. Using knowledge of the dynamics and input controls, the paper proposes an approach to estimate the point spread function (PSF), which is a blur kernel that describes the camera motion during an exposure window. This PSF is then used to remove motion blur and detailed experimental results are presented.

In this paper Cupec, Nyarko, Filko, Kitanov and Petrović address the issue of localization, which they treat as a recognition problem on depth images. The authors explore the use of higher-level geometric primitives, such as 3D planar surfaces and line segments extracted from depth images for place recognition and probabilistic matching, along with the pose estimation. The use of structural geometric information makes the strategy robust with respect to lighting, occlusions and dynamic changes in the environment. The method is evaluated for indoor environments where the geometric primitives are readily extracted and efficiently matched with those in the previously constructed environment model.

The final paper in this special issue considers the problem of finding appropriate robotic grasps. Lenz, Lee and Saxena consider this as a detection problem: given noisy and partial views of objects from an RGB-D sensor viewing a scene containing multiple objects, what are the best locations to place the robot's gripper? It exploits the use of recently popularized deep learning techniques to learn effective features for grasp detection. The authors devote particular care to the design of the network architecture and run-time evaluation, so as to enable fast and robust detection and handling of multimodal data. The method is shown to outperform the previously proposed methods and results are presented for a Baxter robot grasping objects of a table top.

We are excited and encouraged by these latest results, which show great progress in the long-held dream of creating competent and capable robots that are guided and informed significantly by vision sensors, as we are. This special issue presents systems capable of working over large scales and long time periods in complex indoor and outdoor natural environments and with often adverse viewing conditions. We hope you enjoy this issue devoted to fostering links between research in robotics and vision. We look forward to seeing further contributions in this fascinating research area in future special issues on this topic.

Guest editors:
Jana Košecká
Department of Computer Science
George Mason University
Fairfax, Virginia, USA
kosecka@cs.gmu.edu

Eric Marchand
Université de Rennes 1, IRISA
Inria Rennes-Bretagne Atlantique
Campus de Beaulieu, Rennes, France
Eric.Marchand@irisa.fr

Peter Corke
Australian Centre for Robotic Vision
Queensland University of Technology
Australia
peter.corke@roboticvision.org