In: Workshop on RGB-D: Advanced Reasoning with Depth Cameras, Robotics: Science and Systems (R:SS 2014), Berkeley, CA (2014).

Real-time RGB-D-based Object and Manipulator Pose Estimation

Karl Pauwels*, Leonardo Rubio*, Vladimir Ivan†, Sethu Vijayakumar† and Eduardo Ros*

*Computer Architecture and Technology Department, University of Granada, Spain
{kpauwels,lrubio,eros}@ugr.es

†School of Informatics, University of Edinburgh, UK
{v.ivan,sethu.vijayakumar}@ed.ac.uk

***Abstract*—We present an overview of our recent work on real-time model-based object pose estimation from intensity and depth cues. We have developed a system that can simultaneously track the pose of hundreds of rigid objects. By incorporating proprioceptive information, objects can be tracked together with their robotic manipulator, enabling accurate visual servo-control even in the presence of severe camera motion. By imposing constraints on the relative poses of object parts, the same system can be used to detect and track the pose of articulated objects as well.**

I. Introduction

We have developed a model-based object pose detection and tracking system that relies on a continuous real-time interaction between visual *simulation* and visual *perception*, exploiting respectively the graphics and compute capabilities of modern Graphics Processing Units (GPUs). The object poses are updated by combining multiple dense and sparse low-level visual cues (depth, motion, keypoints). In turn, this updated pose information drives an extensive feedback signal (consisting of model shape and appearance, occlusions, object segments, and pose reliability) back to the signal level where it facilitates the cue extraction itself [6]. An overview is provided in Fig. 1A.

Our tracker can simultaneously estimate the full six Degrees Of Freedom (DOF) pose of 150 arbitrarily shaped rigid objects at 40 frames/second [2, 8]. This very large scalability allows for (the real-time use of) a unified model representation that combines redundant DOFs with flexible constraints [8]. This representation can be applied to (multi-camera) tracking a single or multiple rigid objects, as well as to tracking complex articulated objects together with their manipulators [2, 7]. These properties make the method particularly effective for estimating, recognizing, and tracking the state of objects under manipulation.

II. Multi-object Pose Detection and Tracking

Our approach centers on a novel version of the Iterative Closest Point algorithm for signal-level fusion that allows simultaneous use of multiple dense motion and depth cues for object and manipulator pose estimation [6]. There are three sources of feedback in the system: (1) a novel motion cue called Augmented Reality flow that enables drift-free pose tracking by feeding back appearance information from the model, (2) shape feedback from the models, which facilitates dense disparity computation in case a stereo- rather than RGB-D-sensor is used, and finally (3) the image regions occupied by the objects (while accounting for occlusions) which enables segmentation of the low-level cues. The current real-time system supports arbitrarily shaped rigid objects by matching dense depth information to 3D models. The model complexity is not a bottleneck in the system since GPU graphics hardware is used to efficiently render the scene. The tight integration between graphics and compute in our system enables precise spatial feedback of the tracked object (also considering occlusions due to Z-buffering) at the current pose. This segmentation is iteratively improved to maximize the match with the signal-level information.

The current system enables segmenting and assigning 500,000 valid low-level vision samples to 150 objects in order to robustly track the joint 900 DOFs at 40 frames/second (including low-level cue estimation) on a NVIDIA Geforce GTX 590 GPU.

On benchmark sequences that we specifically designed for this problem, we have shown that our system outperforms state-of-the-art methods (edge-based particle filters, region-based, tracking-by-detection) in terms of accuracy, robustness, and speed [6]. We have also demonstrated excellent performance on a wide variety of challenging real-world sequences involving object manipulation and multiple interacting objects (see Fig. 2A,B and the video at [4]).

III. Joint Object and Manipulator Tracking

By incorporating real-time proprioceptive information and kinematic constraints, the approach allows for highly accurate manipulator tracking as well (Fig. 1B). Using a unique two-way consistency resolution paradigm, we exploit the information provided by grasped objects to facilitate manipulator tracking and vice versa.

We have demonstrated the robustness and accuracy of this system on a complex real-world manipulation task involving active endpoint closed-loop visual servo-control in the presence of both camera and target object motion (Fig. 2C–F and the video at [5]). We partially account for the camera motion by tracking the robot base (*i.e.* considering it as an additional rigid object). We have also explicitly demonstrated

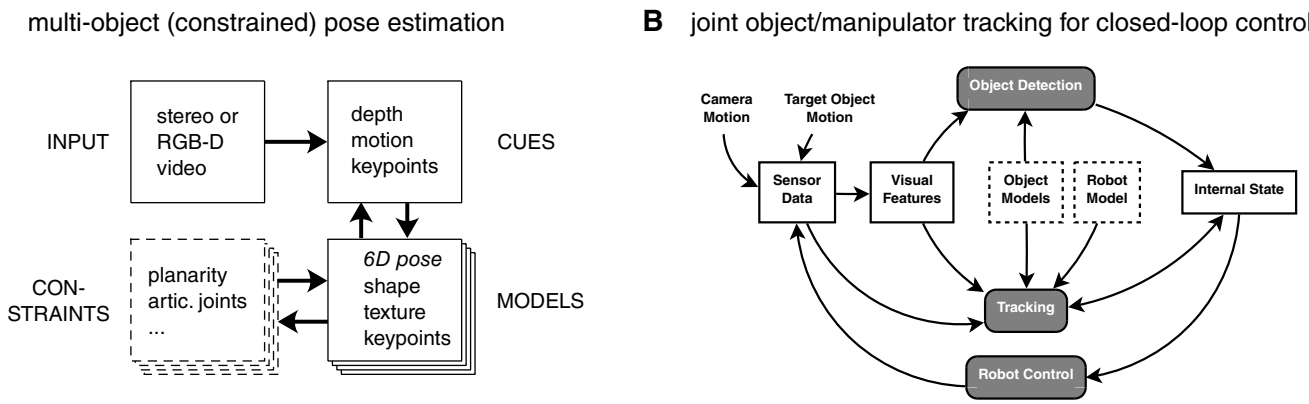

Fig. 1. (A) General overview of multi-object rigid pose estimation, possibly with constrained relative poses and (B) incorporation of manipulator tracking for closed-loop visual servo-control.

the importance of tracking object and manipulator jointly. Ignoring the manipulator (and thus mutual occlusions) results in rapid tracking loss. In many situations there are also not enough measurements at the grasped object (bottle in Fig. 2C) for reliable pose tracking. Tracking only the manipulator results in frequent tracking failures as well, such as in the absence of Kinect depth measurements (Fig. 2D).

IV. Articulated Pose Detection and Tracking

To move from multiple rigid objects to articulated objects, we impose constraints on the relative pose updates between different objects or object parts. For example, in Fig. 2G, the three objects are constrained to move in a plane defined by the objects themselves. In this case, the constraint is violated since the cereal-box is being lifted. For articulated objects, the constraints define the joints. They are imposed after computing the rigid pose updates of the object parts (thus with redundant DOFs) which allows re-use of the highly efficient parallel machinery of Section II.

In addition, we have introduced a novel rigidization framework for optimally handling unobservable parts during tracking [8]. This involves rigidly attaching the minimal amount of unseen parts to the rest of the structure in order to most effectively use the currently available knowledge. We have shown how this framework can be used also for detecting rather than tracking which allows for automatic system initialization or incorporating pose estimates obtained from independent object part detectors. The overhead imposed by the constraint-framework is minimal. For example, the 12-part articulated object of Fig. 2H–J can also be tracked at around 40 frames/second. See the video at [3] for more results.

When simplifying certain aspects of the algorithm (reducing the number of orientations considered by the low-level vision engine, and the number of iterations performed by the pose estimation) this same 12-part object can be tracked on a lightweight Geforce GT 640M LE mobile GPU at 20 frames/second.

V. Discussion

In addition to the visual servoing application mentioned above, the pose estimation system has also been used recently to investigate the combination of visual and acoustic data in detecting human manipulation actions [9] and for identifying the content of a container by fusing tactile and visual feedback in combination with grasping [1]. The rich low-level information and the ability to track a large number of DOFs has the potential to further extend the framework towards deformable models and/or for requiring less precise model definitions.

Acknowledgments

This work has been supported by a Marie Curie Intra European Fellowship (FP7-PEOPLE-2011-IEF-301144). The GPU used for this research was donated by the NVIDIA Corporation.

References

[1] Püren Güler, Yasemin Bekiroglu, Karl Pauwels, and Danica Kragic. What's in the container? Classifying object contents from vision and touch. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Chicago, 2014.

[2] Karl Pauwels. Real-time 3D pose estimation of hundreds of objects. NVIDIA GPU Technology Conference (GTC), 2014.

[3] Karl Pauwels. Real-time articulated object pose detection and tracking. https://www.youtube.com/watch?v=3y2ij9rnI60, 2014. YouTube video.

[4] Karl Pauwels. Real-time pose estimation of hundreds of objects. https://www.youtube.com/watch?v=H4jSU0M7fYc, 2014. YouTube video.

[5] Karl Pauwels. Real-time object pose estimation under imprecise calibration. https://www.youtube.com/watch?v=ijhX8xfXKPE, 2014. YouTube video.

[6] Karl Pauwels, Leonardo Rubio, Javier Díaz Alonso, and Eduardo Ros. Real-time model-based rigid object pose

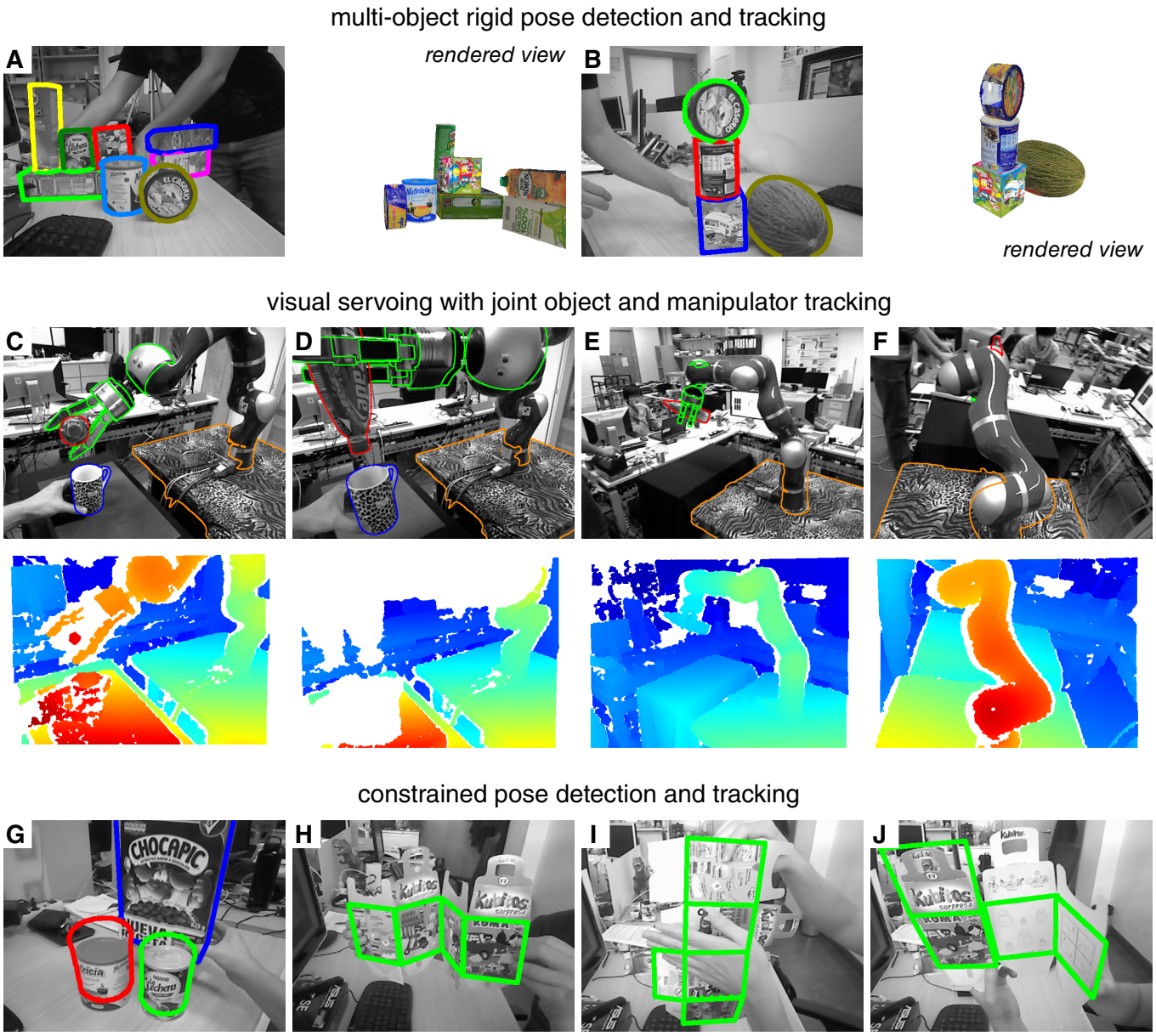

Fig. 2. (A,B) multi-object rigid pose estimation showing segmentation (left) and an arbitrary rendered camera view (right). (C-F) joint object/manipulator tracking showing segmentation (top) and Kinect depth (bottom): (C) relying on the hand to track the bottle and (D) vice versa (in the absence of Kinect depth), robustness to (E) large distance range and (F) severe self-occlusion. (G–J) enforcing constraints between objects and/or object parts: (G) planarity constraint (violated) and (H–J) articulation constraints.

estimation and tracking combining dense and sparse visual cues. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2347–2354, Portland, 2013.

[7] Karl Pauwels, Vladimir Ivan, Eduardo Ros, and Sethu Vijayakumar. Real-time object pose recognition and tracking with an imprecisely calibrated moving RGB-D camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Chicago, 2014.

[8] Karl Pauwels, Leonardo Rubio, and Eduardo Ros. Real-time model-based articulated object pose detection and tracking with variable rigidity constraints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, 2014.

[9] Alessandro Pieropan, Giampiero Salvi, Karl Pauwels, and Hedvig Kjellström. Audio-visual classification and detection of human manipulation actions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Chicago, 2014.