Mechatronics 21 (2011) 983–992

Contents lists available at ScienceDirect

Mechatronics

journal homepage: www.elsevier.com/locate/mechatronics

Interoperable vision component for object detection and 3D pose estimation for modularized robot control

Yasushi Mae *, Jaeil Choi, Hideyasu Takahashi, Kenichi Ohara, Tomohito Takubo, Tatsuo Arai

Department of Systems Innovation, Division of Systems Science and Applied Informatics, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

ARTICLE INFO

Article history:
Received 14 July 2009
Accepted 27 March 2011
Available online 19 April 2011

Keywords:
Robot vision
Object detection
Pose estimation
Local features
Feature matching

ABSTRACT

Finding objects and tracking their poses are essential functions for service robots, in order to manipulate objects and interact with humans. We present novel algorithms for local feature matching for object detection, and 3D pose estimation. Our feature matching algorithm takes advantage of local geometric consistency for better performance, and the new 3D pose estimation algorithm solves the pose in a closed-form using homography, followed by a non-linear optimization step for stability. Advantages of our approach include better performance, minimal prior knowledge for the target pattern, and easy implementation and portability as a modularized software component. We have implemented our approach along with both CPU and GPU-based feature extraction, and built an interoperable component that can be used in any Robot Technology (RT)-based control system. Experiment shows that our approach produces very robust results for the estimated 3D pose, and maintain very low false positive rate. It is also fast enough to be used in on-line applications. We integrated our vision component in an autonomous robot system with a search-and-grasp task, and tested it with several objects that are found in ordinary domestic environment. We present the details of our approach, the design of our modular component design, and the results of the experiments in this paper.

1. Introduction

Object detection and pose estimation is one of essential functions of an autonomous service robot, since we want the robot to be able to find and manipulate objects and interact with humans more intelligently. Although it is possible to make the problem simpler by tagging each object with RFID and other types of equipments and sensors, it is highly desirable to use vision as the primary sensing mode, because we can use vision-based system in more general and broad range of situations without changing the environment due to non-intrusive nature of vision.

Computer vision communities have put tremendous amount of effort on object recognition, and developed so many approaches to finding objects in the scene and recognizing the object's class/category. Their approaches can be roughly classified into two: model-based and appearance based. In model-based approaches, features like edges, straight lines, contours, and segmentations are used to compare the image with the 2D model or the projections of the 3D model of the object. Since the model-based approaches had many disadvantages such as the difficulties in model construction, fitting with generalized model, and occlusions, it became less and less popular since 1990's. Appearance-based approaches also can be divided into many branches depending on the type of features they use: global shape, statistical charateristics, or local features, for example. In particular, local features such as Harris corners [1], KLT [2], Maximally Stable External Regions [3], and SIFT [4] have been used very successfully in object recognition, matching, and tracking, and recent achievements have resolved many issues like robustness and scale/rotation invariance.

When it comes to object detection and pose estimation for robot vision, emphasis is on the accuracy of the estimated 3D pose, because we want the robot to manipulate the object with its robotic hand. To achieve this challenging goal, the state-of-art approaches trys to find the instance(s) of the particular object in the scene, instead of working with a generalized model/pattern. In [5], they compare the overall shape of the object in the scene with 'training views' generated from the 3D model of the objects, assuming the object in the test image is not occluded and can be segmented from the background. This approach might be suitable when the object does not have enough texture and local features, but its assumptions are too strong and the approach is prone to failure in a complex real-world environment. Other researches such as [6–8] combines 3D model of objects with local features. By taking advantage of the robustness and invariance of the corresponding local features, these approaches work well as long as the object have enough texture, even in a complex environment with

* Corresponding author.
E-mail address: mae@sys.es.osaka-u.ac.jp (Y. Mae).

0957-4158/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mechatronics.2011.03.008