Engineering Applications of Artificial Intelligence 63 (2017) 33–44

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Fast scene analysis using vision and artificial intelligence for object prehension by an assistive robot

C. Bousquet-Jette, S. Achiche *, D. Beaini, Y.S. Law-Kam Cio, C. Leblond-Ménard, M. Raison

Department of Mechanical Engineering, Polytechnique de Montréal, C.P. 6079, succ. CV, Montréal, Québec, Canada H3C 3A7

ARTICLE INFO

Article history:
Received 2 August 2016
Received in revised form
7 December 2016
Accepted 20 April 2017

Keywords:
Scene analysis
Artficial vision
Artificial intelligence
Object grasping
Assistive robotics

ABSTRACT

Robotic assistance for people affected by motor deficits is a fast growing field. In this context, two major challenges remain in terms of automated scene analysis and automated object prehension. More specifically, the most robust of current segmentation methods are still computationally intensive, preventing the automation of objects prehension from being fast enough to be considered acceptable as an everyday technical-aid. The objective of this study is to develop a fast scene analysis using vision and artificial intelligence for object prehension by an assistive robot.

The solution developed in this paper aims at facilitating human-machine interaction by enabling users to easily communicate their needs to the technical aid. To achieve this, this paper proposes several novelties in three interconnected domains: scene segmentation, prehension and recognition of 3D objects. A novel technic, inspired by mechanical probing, is developed for scenes probing to detect objects. A simple, fast and effective decision tree is proposed for object prehension. Finally, the physical characteristics of the 3D objects are directly used in the neural network without using discriminants features descriptors.

The results obtained in this paper have shown that scene analysis for robotic object prehension in cooperation with a user can be performed with effective promptness. Indeed, the system requires on average 0.6 s to analyze an object in a scene. With the JACO robotic assistance arm, the system can pick up a requested object in 15 s while moving at 50 mm/s, which may be greatly improved upon using a faster robot. The system performance averages 83% accuracy for object recognition and is able to use a decision tree to select a simple approach path for the robot end-effector towards a desired object. This system, in combination with an assistive robot, has great potential for providing users suffering from musculoskeletal disorders with improved autonomy and independence, and for encouraging sustained usage of this type of technical aids.

1. Introduction

Vision-supported robotic assistance is flourishing, particularly for people affected by age-related loss of mobility and people subject to musculoskeletal disorders. Indeed, robots are often equipped with grippers to improve the autonomy and the capabilities of users who are subject to upper limb motor limitations. Users may be assisted in daily tasks such as manipulating objects, opening doors, and operating electrical switches. One means of controlling the robots is by using a joystick attached to the armrest of a motorized wheelchair. The effectiveness of robot usage in this kind of context was evaluated by (Maheu et al., 2011), which concludes that such robots are very useful. However, the operation of such robots requires good manual dexterity and may therefore be too demanding for certain users, which may cause the object manipulation to be too slow and make complex operations tiring and frustrating. Indeed, the ease of use of this type of robotic aids and the ability to manipulate objects reasonably quickly are essential criteria for their sustained usage. Consequently, it is imperative to design a means of facilitating object prehension and of optimizing the time required for operations that make use of computer vision and artificial intelligence.

In this context, a major current challenge is fast automated scene analysis using vision and artificial intelligence for object prehension by an assistive robot.

Some existing methods are both very promising and advanced, with 3D segmentation being able to handle occluding obstacles in

Abbreviations: OBJ, 3D object file format; PLY, "Polygon File Format"; STL, "Standard Triangle Language" file format

* Corresponding author.
E-mail addresses: christopher.bousquet-jette@polymtl.ca (C. Bousquet-Jette), sofiane.achiche@polymtl.ca (S. Achiche), dominique.beaini@polymtl.ca (D. Beaini), yann-seing.law-kam-cio@polymtl.ca (Y.S. Law-Kam Cio), cedric.leblond-menard@polymtl.ca (C. Leblond-Ménard), maxime.raison@polymtl.ca (M. Raison).

http://dx.doi.org/10.1016/j.engappai.2017.04.015
0952-1976/

complex scenes and to provide shape reconstruction for improved prehension. Some methods develop innovative, complex, functional models for object prehension, and some have obtained good results for object recognition with up to 98.5% performance. Nonetheless, the most robust of the current segmentation methods are still computationally intensive, preventing automated object prehension by robots quickly enough to be acceptable for everyday aid.

For example, the Markov Random Fields method (Anguelov et al., 2005) enables detection and segmentation of complex objects by learning surface and volumetric features to establish a statistical model that allows objects with dissimilar parts to be distinguished, such as a tree, a building, and so on. The Markov Random Fields method does, however, have difficulty distinguishing objects that are similar to one another. (Mian et al., 2005) presents a comparative summary of the algorithms used in the literature (Random Sample Consensus, Graph Matching, Spin Image Matching, Huber's Framework, etc.) for 3D segmentation on the basis of criteria that include the ability to function on free-form objects, computational efficiency, and many others. This research work shows that no current algorithm satisfies the criteria and that none are quick enough. According to the reference, the best current algorithm uses Huber's Framework.

Given that no quick, simple, automated method exists for a fast scene analysis using vision and artificial intelligence for object prehension by an assistive robot, accordingly, the **objective** of this study is to develop such an algorithm. The algorithm must provide answers to all the following questions more quickly than existing methods do: 1. How many objects are there, and where? 2. How should the objects be grasped, i.e., which prehension targets on the objects are effective and what is the favored path of approach for the robot? 3. What are the objects in the scene, as identified by a neural network using data from an active camera?

The research **hypothesis** is that the time required for a robot to grasp an object may be greatly reduced by improving the efficiency of scene analysis and analysis of prehension mode. This hypothesis would be refuted if, in comparison with traditional methods, diminishing the computation times of these analyses by half were not achieved. As to our knowledge, the best analysis time for one scene was 4 s (Lai et al., 2014), therefore the time required should be less than 2 s.

2. Literature review

This section presents the state of the art in objects detection, object prehension, and object recognition.

(Johnson and Hebert, 1999) presented an innovative method known as the spin-image method, which efficiently generates large datasets for 3D objects. This was done by identifying good descriptors of surfaces, for the sake of problems in object detection and in identification of object features, namely, superposition, noise, disorder, and occlusion. After the spin-image compression, the algorithm uses principal component analysis to determine the most significant features of the images and thereby lowers the order of complexity. However, the computation time with this technique is long and highly variable depending on the analyzed scene.

(Mian et al., 2006) presented a more efficient way of obtaining multiple views of 3D objects to generate a large dataset from real and synthetic models. The library contains a table of 3D models with multiple views. The authors report 95% of recognition, and unlike the spin-image method, their algorithm is insensitive to the dimensions of library models. They obtained a computation time of 6 min compared to the 480 min needed for spin-image using a similar computer and scenes. The algorithm still requires at least 2 min for a complex scene, which we consider too long for our application.

(Rusu et al., 2009) proposed a method for segmenting a cloud of 3D points located on a horizontal support surface, such as a table in a domestic environment. That reference estimated the primitive shapes of an object: spheres, cylinders, cones, and prisms. The reconstruction of 3D objects in terms of primitive shapes allowed occluded or otherwise non-visible parts to be inferred during acquisition to facilitate precision and robotic prehension. Designing a 3D object database to aid detection, segmentation, and recognition helped contend with non-visible parts of objects, but time-sensitive techniques suffer greatly from database searches.

(Rusu et al., 2010) included a descriptor for 3D features called Viewpoint Feature Histogram, which works with data sets that are noisy and lack depth information. This was done by detecting object geometry and pose using stereovision. The authors report 98.5% object recognition. They, however, do not report the computational time.

For prehension (Lippiello et al., 2013) presented a method for elastic visual reconstruction of prehension surfaces which ran until the reconstruction fitted the object envelope. Furthermore, the authors proposed a prehension algorithm that ran in parallel with the scene analysis to minimize the total time. Mathematically, the method modeled the points as masses linked by springs and dampers, so as to marry the points to the object envelope. The algorithm moved the robot fingers based on the end-effector kinematic configuration and the weight distribution of the grasped object. This method was completely adaptive to objects and took end-effector kinematic configuration into consideration. However, this method used a complex and time-consuming analytical model, which seems excessive to transfer to our robotic application.

Further, for recognition, ImageNet is a dataset including more than 1.2 million high-resolution images categorized into 1000 classes by a deep convolutional neural network. (Krizhevsky et al., 2012) showed that such a network was able to exceed classification records using only supervised learning. It should be remarked that modifying the network structure by removing convolutions degrades the classification results. That reference illustrated the fact that neural networks are well adapted to recognition of certain patterns following intensive supervised learning, generalizing nonlinear models with decision-making capabilities. This reference also explained that the structure of the neural connections is an important aspect of the training process.

(Lai et al., 2011) presented a hierarchical tree of 51 object classes acquired in RGB-D with the Kinect V1. They showed that combined color and depth information contribute significantly to obtaining good recognition results. This is very relevant to our case. They generated their dataset using a turntable, removing the background and applying the Random Sample Consensus method to identify the surface of the turntable. Then, they employed the spin-image algorithm and the Scale-Invariant Feature Transform descriptor to obtain the image features. Following that, they applied the Linear Support Vector Machine, Gaussian Kernel, and Random Forest methods to classify the objects. They obtained 90.5% recognition in 10 s, which we deem too lengthy for our purposes.

(Tang et al., 2012) presents a system incorporating 3D acquisition, segmentation of objects on a table, meshing, creation of a histogram of features extracted using the Scale-Invariant Feature Transform, and a pose validation. They obtained recognition results in the order of 90% for complex scenes. However, their series of object analyses requires 20 s per scene using a hexa core, 3.2 GHz i7 processor and 24 GB of RAM, which is definitely not fast enough for the needs of our project.

Finally, an object recognition algorithm by unsupervised