International Journal on Recent Researches in Science, Engineering & Technology (IJRRSET)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. Journal Indexed in JIR, DIIF and SJIF.
Available online at: www.jrrset.com

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
Volume 8, Issue 1
January 2020.
JIR IF : 2.54
SJIF IF : 4.334
Cosmos: 5.395

Received on: 04.12.2019 Published on: 20.01.2020 Pages : 24-29

# OBJECT IDENTIFICATION AND RECOGNIZATION TECHNIQUES IN ROBOTIC VISION

R. Nandhakumar[a*], R. Kalyanasundaram [b], M.Roberts Masillamani[c]

[a,b] Department of Mechatronics Engineering, Agni College of Technology, Chennai 600130, Tamilnadu, India
[c] Department of Computer Science and Engineering, Agni College of Tech, Chennai 600130, Tamilnadu, India

[a] e-mail: nandhakumar.mht@act.edu.in, [b] e-mail: kalyanasundaram.mht@act.edu.in

## ABSTRACT

Technological advances are currently being directed to assist the human population in performing ordinary tasks in everyday settings. In this context, a key issue is the interaction with objects of varying size, shape and degree of mobility. Consequently, autonomous assistive robots must be provided with the ability to process visual data in real time so that they can react adequately for quickly adapting to changes in the environment. Reliable object detection and recognition is usually a necessary early step to achieve this goal. In spite of significant research achievements, this issue still remains a challenge when real-life scenarios are considered. In this paper, we present a vision system for assistive robots that is able to detect and recognise objects from a visual input in ordinary environments in real time. The system computes colour, motion and shape cues combining them in a probabilistic manner to accurately achieve object detection and recognition, taking some inspiration from vision science. In addition, with the purpose of processing the input visual data in real-time, a Graphical Processing Unit (GPU) has been employed. The presented approach has been implemented and evaluated on a humanoid robot torso located at realistic scenarios.

## INTRODUCTION

Nowadays robots have found their way from sealed working stations in factories to people's living and working spaces, where they should be able to autonomously perform different services useful to the well-being of humans, such as domestic tasks, healthcare services, entertainment, and education. In particular, with the purpose of improving people's quality of life, especially for the elderly, the field of assistive robotics is becoming increasingly popular. Research is progressing from special-purpose service robots such as autonomous cleaning or transport systems, to multi-functional assistive robots able to integrate diverse abilities such as person detection and tracking, human-robot interaction, reasoning, localization, navigation, object detection and recognition, planning and manipulation. In addition, these assistive robots are expected to operate in a flexible manner, without constraining the environment, and in a reasonable time, while guaranteeing the safety of all their surrounding elements, especially when they are human beings [1] [2]. However, despite the wide research in this area (e.g. Johnny [3], HOBBIT [4], KSERA [5], Cogniron [6], CareO-Bot [7], HERB [8], Accompany [9], AAL4ALL [10] and many others), the progress in assistive robotics has been relatively slow to date. This is mainly due to the fact that the environments to cope with are dynamic, unpredictable and human-oriented.

In addition, depending on the application, long human-robot interactions could miserably fail because of the limited system's autonomy and abilities, as broadly analysed in [11]. Thus, an assistive robot should be provided with a vast set of perception and action capabilities to efficiently perform its goal tasks in real scenarios, while properly interacting with its users along its life. Among all these capabilities, this paper is focused on perception for object detection and recognition, a key task for a meaningful assistance. In this context, vision is considered a primary cue because of the information it can provide. Actually, vision has been used in numerous robotic applications to successfully achieve a task (e.g. obstacle avoidance for navigation [12], [13], [14], [15], human recognition for Human-Robot Interaction [16], [17], activity recognition for cooperative behaviour [18], [19], [20] and object identification for manipulation [21], [22], [23], to name only a few). However, despite significant achievements, the problem of detecting and recognising objects efficiently and accurately still remains a scientific challenge when real scenes are considered. Apart from a great number of objects in the images, the reasons for this difficulty are to be found in issues such as their interactions and occlusions, along with photometric and geometric variations in pose, size, etc. Furthermore, noise in images, the nature of objects themselves, complex object shapes and illumination changes, make it a hard task.

## SYSTEM DESCRIPTION

From a biological point of view, humans are able to easily identify the objects present in their environment. Therefore, insights from human visual processing could be a starting point for developing computer models. This is the case of AlAbsi and Abdullah [24], who designed BIORecS emulating the human vision system. Concretely, BIORecS achieves accurate object recognition in complex scenarios by combining functions of some areas of the human visual cortex and the connection mechanisms between the visual areas in humans, implemented by feedforward and feedback techniques. This model consists of four stages closely intertwined: feature extraction (object shapes are obtained by combining the image edges extracted with Gabor filters); visual attention (a support vector machine is used as object shape classifier); recognition (carried out by Principal Components Analysis) and image database (containing the objects to be recognised). However, although this architecture may allow the system to overcome some key issues in object recognition - such as changes in illumination, occlusions and high-cluttered scenesthe description of objects is not adequate since different objects can have the same visual shape. For example, a ball, a bracelet, a disk, a coin or a drum would all belong to the category of circular shape. Furthermore, some factors such as its pose, scene background or illumination conditions may modify the object's shape. Consequently, a model reformulation is necessary. Alternatively, object detection and recognition could be considered as an attentional mechanism since it refers to the extraction of target information from the observed scene. In this sense, a dorsal attention system could fit. Generally speaking, this system could be defined as a top-down (goaloriented) modulation of stimulus-driven (e.g. saliency) attentional capture by targets versus distractors. In this regard, a four-module attentional architecture has been defined by Lanillos et al. in which the first module corresponds to the perception sense by building an egocentric map according to relevance encoded as saliencyFocusing on the task at hand, the developed visual system should be provided with a perception module which builds a saliency map based on the most distinctive visual features, followed by a module in charge of object recognition. In this way, the system will be centred in the potential targets by reducing the sensory data to be processed and, therefore, making tractable the unmanageable amount of information received from the visual sensors. In addition, a memory that stores information about the objects to be recognised should also be integrated. Therefore, our vision system consists of three different modules (Figure 1):

- Feature Extraction, that generates a saliency map from image segmentation based on three object properties: colour, shape and motion
- Memory, which stores the models of the potential target objects
- Recognition, that is responsible for recognising the objects from the visual input and the data coming from the previous modules

Thus, this architecture is based on a richer object description for robustly detecting and recognising any object in real scenarios without establishing any constraint about the objects and the environment.

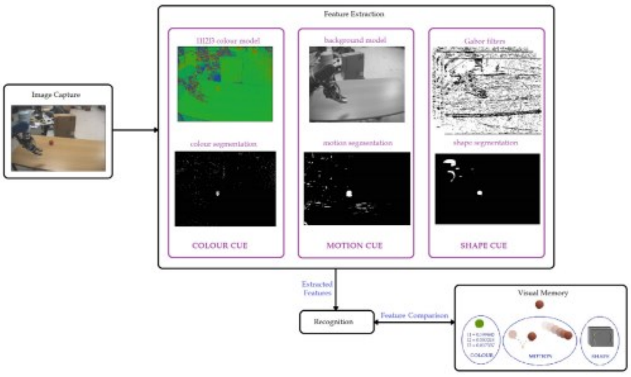

Fig. 1. Overview of the system architecture

## IMPLEMENTATION DETAILS

Real-time processing is a critical demand when state-ofthe-art robot systems are designed. This requirement calls for an efficient processing unit. A solution is to process visual input with a Graphical Processing Unit (GPU), potentially reducing time consumption in a drastic way. However, despite its highly parallel computation capabilities, writing efficient GPU programs is not evident, especially for uneven workloads (e.g. the higher the number of interest objects is, the higher the computational costs are). In particular, our algorithms have been implemented on an NVIDIA GeForce GTX 745. It includes 384 Compute Unified Device Architecture (CUDA) cores with 4-GB memory and chip-level power enhancements. A fast access to shared and GPU's main memories characterizes these CUDA cores. Moreover, graphics API functions are not required for parallel implementations in C language; this is very convenient for properly implementing the necessary parallel algorithms that deal with irregular workloads. The CPU captures an image and uploads it to the GPU, which will perform the subsequent image processing steps, namely, from feature extraction to object recognition. The GPU will return the output to the CPU for it to decide the next action to be performed by the robot. Then, the visual processing starts again. Since object feature detection and tracking is a computationally intensive task, but highly parallelizable, a good parallel solution can be devised to the effect that all image processing is carried out by the GPU (using 1023 threads per block). As a final system output, the CPU shows on the screen the detected objects.

## EXPERIMENTAL RESULTS

The proposed approach for object detection and recognition in real scenarios has been tested in three different kinds of scenarios. First of all, a semi-structured scene was considered so that a methodical study of the efficiency based on different factors could be carried out (e.g. occlusions, light reflexes, changes in illumination, shadows, etc.). Then, the second set of experiments involved two real, cluttered environments in which the target objects were to be found amongst a set of ordinary items such as calendars, books, clocks or pens. Finally, an image dataset has been used to evaluate the performance of the system by means of object instance recognition and in comparison with other state-of-the-art approaches. To conclude, a performance analysis in terms of execution time is presented. For the two first experiments, a humanoid torso endowed with a Robosoft TO40 pan-tilt-vergence stereo head and two multi-joint arms was used. The head mounts two Imaging Source DFK 31BF03-Z2 cameras acquiring colour images at 30 Hz with a resolution of 1024x768 pixels. The baseline between cameras is 270 mm and the motor positions are provided by high-resolution optical encoders.

**Experiment 1:** Semi-structured scenes In the case of semi-structured scenes, the robot was located in front of a table on which the objects were placed. In this experimental setup, the table was initially empty and, after a little while, a human was placing and removing the different objects on the table without interacting directly with the robot system. In this way, the motion cue was instrumental in detecting both the human presence in the robot workspace as well as the new object instance on the table. Actually, in this experiment, the three visual cues have the same weight when the segmentation result is determined. Four different objects have been used as targets: a red ball, a toy car, a bottle and a money box. The object position and orientation were modified for each frame. Obviously, the number of resulting orientations varies based on the considered object; for instance, the red ball has only one orientation, while the toy car was observed in 12 different orientations (approximately every 30 degrees)

**Experiment 2:** Real scenarios In this experiment, the objects to be detected and recognised were placed on a desk. Two unstructured environments were used composed of everyday objects of different nature and features such as textured books, pens, clock, etc. In this context, the objects to be detected and recognised include a red ball, a toy car, a yellow ball, a green bulb, a stapler, and a wooden generalized cylinder. These objects were located at different positions and/or orientations within the considered scenario, resulting partially occluded in some cases. As in the previous case, a human is continuously interacting with the target objects, but not with the robot system, so that the motion cue triggers again a visual attention focus. However, the other two visual cues are required to distinguish between the target objects and other moving elements in the scene such as the person. For this reason, the three cues have the same weight in the object recognition process

**Experiment 3:** Image Repository For the third validation experiment we compare the performance of our approach with state-of-the-art methods by using a public image repository. Actually, given that the ability to recognise objects is crucial for many applications, a wide range of public image repositories is available. These datasets allow researchers to evaluate their approaches with a large number of objects and under different conditions, as well as to compare their performance with other stateof-the-art approaches. However, these repositories could be classified based on the goal to be satisfied. That is, object recognition has multiple levels of semantics (e.g. category recognition, instance recognition, pose recognition, etc.), it can refer to different application scenarios or it could be based on certain input data. Consequently, the required evaluation dataset must correspond to the needs of a particular approach. This is why the RGB-D Object Dataset, publicly available at http://www.cs.washington.edu/rgbd-dataset, has been used for this validation. This dataset is composed of thousands of images of 300 objects commonly found in home and office environments, taken from multiple views by using an RGBD camera

**Experiment 4:** Execution Time Analysis The last evaluation experiment refers to the analysis of the benefits of using the GPU for parallel computing. A similar study was presented by Ferreira et al. in the context of Bayesian models for multimodal perception. With that aim, we carried out a comparison between the performance using parallel and non-parallel computing depending on the image resolution and the number of potential targets.

## CONCLUSION

During the last decades, robotics research moved from stationary robotic systems in constrained environments to mobile and service-oriented robots operating in realistic and unconstrained environments. One rising application field is assistive robotics, aimed at developing robots that support humans as their daily-life assistants. With that aim, these systems must be endowed with different abilities such as localization, mapping, path planning, obstacle avoidance, object detection, recognition, and manipulation. In particular, in this paper we have focused on object detection and recognition. Even though this issue is the heart of different robotic assistive abilities, real-time efficient object detection and recognition is still a challenging problem when real scenarios are considered. Part of this problem is due to the presence of cluttered, dynamic backgrounds, with possible occlusions, interactions and additional photometric and geometric variations. The proposed approach has been implemented on a robotic platform and tested by considering different parameters which might make the system fail. This large number of parameters allows us to analyse the robustness of the proposed method. For further experimental validation, a public image repository for object recognition has been used, allowing a quantitative comparison with respect to other state-of-the-art techniques when real-world scenes are considered. Finally, a temporal analysis of the performance was provided with respect to image resolution and number of target objects in the scene.

## REFERENCES

[1] E. Martinez and A. del Pobil, “Visual surveillance for human-robot interaction,” in SMC, 2012, pp. 3333–3338.

[2] E. Martinez and A. del Pobil, “Safety for human-robot interaction in dynamic environments,” in ISAM, 2009, pp. 327–332.

[3] T. Breuer, G. G. Macedo, R. Hartanto, N. Hochgeschwender, D. Holz, F. Hegger, Z. Jin, C. Muller, J. Paulus, M. Reckhaus, J. A. Ruiz, ¨ P. Ploger, and G. Kraetzschmar, “Johnny: An autonomous service robot ¨ for domestic environments,” J Intelligent and Robotic Systems, vol. 66, pp. 245–272, 2012.

[4] (2013) Hobbit-the mutual care robot. [Online]. Available: http: //hobbit.acin.tuwien.ac.at/

[5] (2013) Ksera-knowledgeable service robots for aging. [Online]. Available: http://ksera.ieis.tue.nl/

[6] (2007) Cogniron-the cognitive robot companion. [Online]. Available: http://www.cogniron.org/final/Home.php

[7] (2015) Care-o-bot. [Online]. Available: http://www.care-o-bot-4.de/

[8] (2015) Herb. [Online]. Available: http://www.cmu.edu/herb-robot/

[9] (2015) Accompany. [Online]. Available: http://www.accompanyproject. eu/

[10] A. Costa, P. Novais, and R. Simoes, “A caregiver support platform within the scope of an ambient assisted living ecosystem,” Sensors, vol. 14, pp. 5654–5676, 2014.

[11] I. Leite, C. Martinho, and A. Paiva, “Social robots for long-term interaction: A survey,” Intl Journal of Social Robotics, vol. 5, pp. 291– 308, 2013.

[12] F. Bonin-Font, A. Ortiz, and G. Oliver, “Visual navigation for mobile robots: A survey,” J Intelligent and Robotic Systems, vol. 53, pp. 263– 296, 2008.

[13] J. Antich, A. Ortiz, and G. Oliver, “A control strategy for fast obstacle avoidance in troublesome scenarios: application in underwater cable tracking,” in IFAC Conf on Manoeuvring and Control of Marine Craft, 2006.

[14] H. Morita, M. Hild, J. Miura, and Y. Shirai, “Panoramic view-based navigation in outdoor environments based on support vector learning,” in IROS, 2006, pp. 2302–2307.

[15] J. Shen and H. Hu, “Visual navigation of a museum guide robot,” in WCICA, vol. 2, 2006, pp. 9169–9173.

[16] D.-H. Lee and J.-H. Kim, “A framework for an interactive robot-based tutoring system and its application to ball-passing training,” in ROBIO, 2010, pp. 573–578.

[17] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, “Machine recognition of human activities: A survey,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 18, pp. 1473–1488, 2008.

[18] O. Chang, “Evolving cooperative neural agents for controlling vision guided mobile robots,” in Intl Conf on Cybernetic Intelligent Systems, 2010, pp. 1–6.

[19] M. Asada, E. Uchibe, and K. Hosoda, “Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development,” Artificial Intelligence, vol. 110, pp. 275–292, 1999.

[20] Y. Kuniyoshi, J. Rickki, M. Ishii, S. Rougeaux, N. Kita, S. Sakane, and M. Kakikura, “Vision-based behaviors for multi-robot cooperation,” in IROS, vol. 2, 1994, pp. 925–932.

[21] D. Kragic and H. Christensen, “Survey on visual servoing for manipulation,” Computational Vision and Active Perception Laboratory, Tech. Rep., 2002.

[22] L. Whitcomb, D. Yoerger, H. Singh, and D. Mindell, “Towards precision robotic maneuvering. survey, and manipulation in unstructured undersea environments,” in Intl Symp on Robotics Research. Springer-Verlag Publications, 1998, pp. 45–54.

[23] Recognizing Patterns in Signals, Speech, Images and Videos. ICPR 2010 Contests, ser. LNCS. Springer Berlin Heidelberg, 2010, vol. 6388.

[24] S. Lee, S. Lee, J. Lee, D. Moon, E. Kim, and J. Seo, “Robust recognition and pose estimation of 3d objects based on evidence fusion in a sequence of images,” in ICRA, 2007, pp. 3773–3779.

[25] N. Sian, T. Sakaguchi, K. Yokoi, Y. Kawai, and K. Maruyama, “Operating humanoid robots in human environments,” in RSS Workshop: Manipulation for Human Environments, 2006.

[26] R. Platt, R. Burridge, M. Diftler, J. Graf, M. Goza, and E. Huber, “Humanoid mobile manipulation using controller refinement,” in RSS Workshop: Manipulation for Human Environments, 2006.

[27] C. Urdiales, M. Dominguez, C. de Trazegnies, and F. Sandoval, “A new pyramid-based color image representation for visual localization,” Image and Vision Computing, vol. 28, no. 1, pp. 78–91, 2010.

[28] C. Zhang, Y. Qiao, E. Fallon, and C. Xu, “An improved camshift algorithm for target tracking in video surveillance,” in 9th IT & T Conf, 2009.

[29] Z. Kim, “Real time object tracking based on dynamic feature grouping with background subtraction,” CVPR, vol. 0, pp. 1–8, 2008.

[30] M. Villamizar, A. Sanfeliu, and J. Andrade-Cetto, “Computation of rotation local invariant features using the integral image for real time object detection,” in ICPR, vol. 4, 2006, pp. 81–85