An Integrated, Modular Framework for Computer Vision and Cognitive Robotics Research (icVision)

Jürgen Leitner, Simon Harding, Mikhail Frank, Alexander Förster and Jürgen Schmidhuber

Abstract We present an easy-to-use, modular framework for performing computer vision related tasks in support of cognitive robotics research on the *iCub* humanoid robot. The aim of this biologically inspired, bottom-up architecture is to facilitate research towards visual perception and cognition processes, especially their influence on robotic object manipulation and environment interaction. The *icVision* framework described provides capabilities for detection of objects in the 2D image plane and locate those objects in 3D space to facilitate the creation of a world model.

1 Introduction

Vision and the visual system are the focus of much research in psychology, cognitive science, neuroscience and biology. A major issue in visual perception is that what individuals 'see' is not just a simple translation of input stimuli (compare *optical illusions*). The research of Marr in the 1970s led to a theory of vision using different levels of abstraction [11]. He described human vision as processing inputs, stemming from a two-dimensional visual array (on the retina), to build a three-dimensional description of the world as output. For this he defines three levels: a 2D (or primal) sketch of the scene (using feature extraction), a sketch of the scene using textures to provide more information, and finally a 3D model.

Visual perception is of critical importance, as the sensory feedback allows to make decisions, trigger certain behaviours, and adapt these to the current situation. This is not just the case for humans, but also for autonomous robots. The visual feedback enables robots to build up a cognitive mapping between sensory inputs

Jürgen Leitner, Mikhail Frank, Alexander Förster and Jürgen Schmidhuber
Dalle Molle Institute for Artificial Intelligence (IDSIA), USI/SUPSI, Lugano, Switzerland
e-mail: juxi@idsia.ch

Simon Harding
Machine Intelligence, Ltd UK, e-mail: simon@machineintelligence.co.uk

and action outputs, therefore closing the sensorimotor loop. Thus being able to perform actions and adapt to dynamic environments. We are aiming to build a visual perception system for robots, based on human vision, that allows to provide this feedback leading to more autonomous and adaptive behaviours.

Our research platform is the open-system humanoid robot *iCub* [17] developed within the EU funded ‘RobotCub’ project. In our setup, as shown in Figure 1 (left), it consists of two anthropomorphic arms, a head and a torso and is roughly the size of a human child. The *iCub* was designed for object manipulation research. It also is an excellent experimental, high degree-of-freedom (DOF) platform for artificial (and human) cognition research and embodied artificial intelligence (AI) development [13]. To localise objects in the environment the *iCub* has to rely solely, similarly to human perception, on a visual system based on stereo vision. The two cameras are mounted in the head. Their pan and tilt can jointly be controlled, with vergence providing a third DOF. The neck provides 3 more DOF for gazing.

We describe a framework, named *icVision*, supporting the learning of hand-eye coordination and object manipulation, by solving visual perception issues in a biologically-inspired way.

2 The *icVision* Framework

Research on perception has been an active component of developing artificial vision (or computer vision) systems, in industry and robotics. Our humanoid robot should be able, like the human mind, learn to perceive objects and develop a representation that allows it to detect this object again. The goal is to enable adaptive, autonomous behaviours based on visual feedback by combining robot learning approaches (AI and machine learning (ML) techniques), with computer vision.

This framework was developed to build a biologically-inspired architecture (in-line with the description by Marr). It processes the visual inputs received by the cameras and builds (internal) representations of objects. It facilitates the 3D localisation of the detected objects in the 2D image plane and provides this information

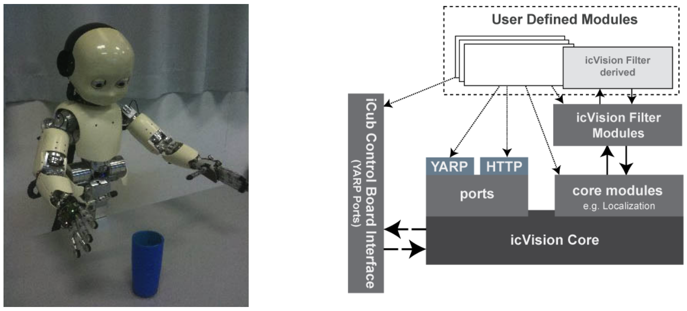

Fig. 1 Left: The *iCub* humanoid robot. **Right:** Architecture of the *icVision* framework.

to other systems (e.g. motion planner). Figure 1 (right) sketches the *icVision* architecture. The system consists of distributed YARP modules[1] interacting with the *iCub* hardware and each other. The specialised modules can be connected and form pathways to perform, for example, object detection, similarly to the hierarchies in human perception in the visual cortex (V1, V2, V3, ...) [6].

The main module, the *icVision* **Core**, handles the connection with the hardware and provides housekeeping functionality (e.g., GUI, module start/stop). Implemented modules include object detection (aka filters), 3D localisation and a gaze controller interface based on the position of the object in the image plane (as provided by the filters). These are reachable via standardised interfaces allowing for easy swapping and reuse of modules and extending functionality. For example, other brain-inspired modules, a saliency & a disparity map, have recently been added.

*2.1 Detecting Objects (***icVision** *Filter)*

The first thing done by the human visual system, and investigated by us, is the segmentation (detection) in the visual space (the 2D images). There exists a vast body of work on all aspects of image processing [3], using both classical and machine learning approaches. The *icVision* filter modules, which relate to Marr's first and second level, provide object detection in the images. As input the filter module provides the camera image in grayscale and split into RGB and HSV channels. The result of the filter is a binary segmentation of the camera image for a specific object. Figure 2 shows a tea box being tracked by the *iCub* in real-time using a learned filter. Also in Figure 3 the binary output can be seen.

Using machine learning, more complicated filters can be generated automatically instead of engineered. We apply Cartesian Genetic Programming (CGP) [14, 15] to provide automatic generation of computer programs making use of the functionality integrated in the OpenCV image processing library [1], therefore incorporating domain knowledge. It provides an effective method to learn new object detection algorithms that are robust, if the training set is chosen correctly [4].

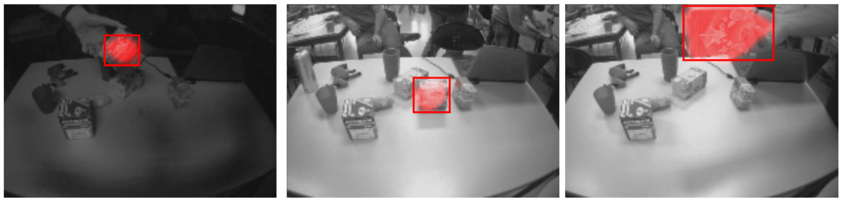

Fig. 2 The detection of a tea box in changing lighting condition performed by a learned filter. The binary output of the filter is used as red overlay.

[1] YARP [12] is a middleware that allows easy, distributed access to the sensors and actuators of the *iCub* humanoid robot, as well as, to other software modules.

*2.2 Locating Objects (***icVision** *3D)*

To enable the robot to interact with the environment it is important to localise the object first. Developing an approach to perform robust localisation to be deployed on a real humanoid robot is necessary to provide the necessary inputs for on-line motion planning, reaching, and object manipulation.

The *icVision* 3D localisation module is one of the modules provided by the core framework. It allows for conversion between camera image coordinates and 3D coordinates in the robot reference frame. Using the objects location in the cameras (provided by an *icVision* Filter module) and pose information from the hardware, this module calculates where the object is in the world. This information is then used to update the world model. Figure 3 describes the full 3D location estimation process, starting with the camera images received from the robot and ending with the localised object being placed in our MoBeE world model [2].

Stereo Vision describes the extraction of 3D information out of digital images and is similar to the biological process of stereopsis in humans. Its basic principle is the comparison of images taken of the same scene from different viewpoints. To obtain a distance measure the relative displacement of a pixel between the two images is used [5]. While these approaches, based on projective geometry, have been proven effective under carefully controlled experimental circumstances, they are not easily transferred to robotics applications. For the *iCub* platform several approaches have previously been developed. The 'Cartesian controller module' [16], for example, provides basic 3D position estimation functionality and gaze control. This module works well on the simulated *iCub*, however it is not fully supported and functional on the hardware platform, and therefore does not perform well. The most accurate currently available localisation module for the *iCub* exists in the 'stereoVision' module. It provides accuracy in the range of a few centimetres.

The *icVision* 3D localisation module provides an easy way of swapping between various localisation implementations, including the two mentioned. We also provide an implementation estimating the location using machine learning [8, 10].

3 Use Cases and Current Application of the Framework

This framework has already successfully been used in our research. Here we give a short list of use cases for the *icVision* framework.

The full system has been used together with a reactive controller to enable the *iCub* to autonomously re-plan a motion to avoid an object it sees [2]. The object is placed into the world model purely from vision, it is able to update the position of the object in real-time, even while the robot is moving.

The learning of specific filters for certain objects was done using CGP (as mentioned above) [4, 9]. To allow for a more autonomous acquisition of object representations *icVision* filters are learned from objects perceived by the cameras. By using a saliency map and standard feature detectors, we were able to provide the needed

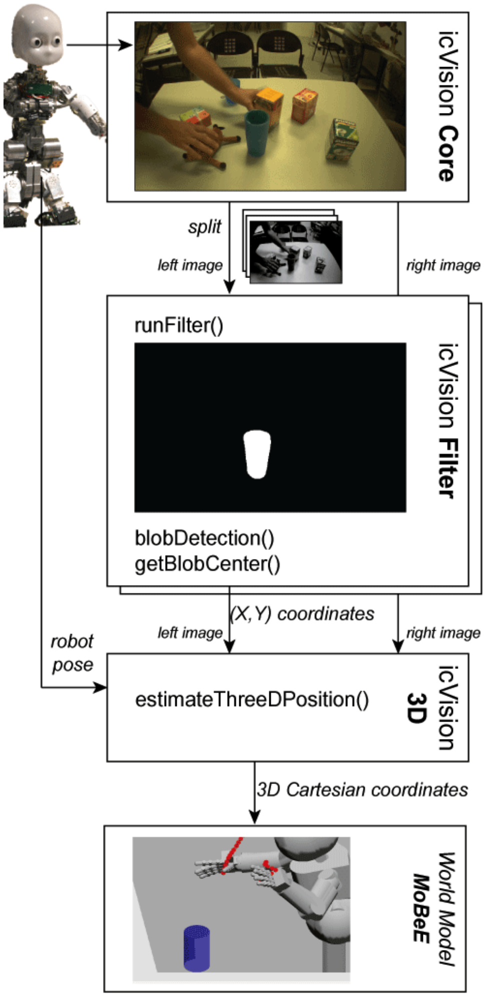

Fig. 3 The 3D location estimation works the following: At first the camera images are acquired from the hardware via YARP. The images are converted into grayscale, as well as, split into RGB/HSV channels and distributed to all active *icVision* filters.
Each filter then processes the images received using OpenCV functions. (ending with a thresholding operation). The output of this is a binary image, segmenting the object to be localised.
A blob detection algorithm is run on these binary images to find the (centre) location of the detected object in the image frame.
The position of the object in both the right and left camera images is sent to the 3D localisation module, where together with the robots pose, i.e. the joint encoders, a 3D location estimation is generated.
As the last step the localised object is then placed in the existing world model.

inputs to our CGP learner for building robust filters [7]. We are in the process of learning filters for the robot's fingers to perform research in how the humanoid can develop sensorimotor control.

To add our 3D localisation approach to the framework, we used a Katana robotic arm to teach the *iCub* how to perceive the location of the objects it sees. The Katana positions an object within the shared workspace, and informs the *iCub* about the location. The *iCub* then moves to observe the object from various angles and poses. Its pose and the 2D position outputs provided by an *icVision* filter are used to train artificial neural networks (ANN) to estimate the object's Cartesian location. We show that satisfactory results can be obtained for localisation [10]. Furthermore, we demonstrate that this task can be accomplished safely using collision avoidance software to prevent collisions between multiple robots in the same workspace [2].

4 Conclusions

We combine the current machine learning and computer vision research to build a biologically-inspired, cognitive framework for the *iCub* humanoid robot. The developed *icVision* framework facilitates the autonomous development of new robot controllers. Cognition and perception are seen as the foundation to developmental mechanisms, such as as sensorimotor coordination, intrinsic motivation and hierarchical learning, which are investigated on the robotic platform.

The reason for the focus on vision is twofold, firstly the limited sensing capabilities of the robotic platform and secondly, vision is the most important sense for humans. As we use a humanoid robot investigating how humans do this tasks of perception, detection and tracking of objects is of interest. These facilitate the building of a world model, which is used for tasks like motion planning and grasping. Realtime, incremental learning is applied to further improve perception and the model of the environment and the robot itself. Learning to grasp and basic hand-eye coordination are the areas of research this framework is currently applied.

References

1. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
2. Frank, M., et al.: The modular behavioral environment for humanoids and other robots (mobee). In: Int'l. Conference on Informatics in Control, Automation and Robotics (2012)
3. Gonzalez, R., Richard, E.W.: Digital image processing (2002)
4. Harding, S., Leitner, J., Schmidhuber, J.: Cartesian genetic programming for image processing. In: Genetic Programming Theory and Practice X (to appear). Springer (2012)
5. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge University Press (2000)
6. Hubel, D., Wensveen, J., Wick, B.: Eye, brain, and vision. Scientific American Library (1995)
7. Leitner, J., et al.: Autonomous learning of robust visual object detection on a humanoid (2012). (submitted to IEEE Int'l. Conference on Developmental Learning and Epigenetic Robotics)
8. Leitner, J., et al.: Learning spatial object localisation from vision on a humanoid robot. (submitted to International Journal of Advanced Robotics Systems) (2012)
9. Leitner, J., et al.: Mars terrain image classification using cartesian genetic programming. In: International Symposium on Artificial Intelligence, Robotics & Automation in Space (2012)
10. Leitner, J., et al.: Transferring spatial perception between robots operating in a shared workspace. (submitted to IROS) (2012)
11. Marr, D.: Vision: A Computational Approach. Freeman & Co., San Francisco (1982)
12. Metta, G., Fitzpatrick, P., Natale, L.: YARP: Yet Another Robot Platform. International Journal of Advanced Robotics Systems **3**(1) (2006)
13. Metta, G., et al.: The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks **23**(8-9), 1125–1134 (2010)
14. Miller, J.: An empirical study of the efficiency of learning boolean functions using a cartesian genetic programming approach. In: Genetic and Evolutionary Computation Conference (1999)
15. Miller, J.: Cartesian genetic programming. Cartesian Genetic Programming pp. 17–34 (2011)
16. Pattacini, U.: Modular Cartesian Controllers for Humanoid Robots: Design and Implementation on the iCub. Ph.D. thesis, RBCS, Italian Institute of Technology, Genova (2011)
17. Tsagarakis, N.G., et al.: iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research. Advanced Robotics **21**, 1151–1175 (2007)