A Portable Active Binocular Robot Vision Architecture for Scene Exploration

Aamir Khan, Gerardo Aragon-Camarasa, and J. Paul Siebert

School of Computing Science, University of Glasgow, G12 8QQ, Glasgow, UK
Gerardo.AragonCamarasa@glasgow.ac.uk

Abstract. We present a portable active binocular robot vision architecture that integrates a number of visual behaviours. This vision architecture inherits the abilities of vergence, localisation, recognition and simultaneous identification of multiple target object instances. To demonstrate the portability of our vision architecture, we carry out qualitative and comparative analysis under two different hardware robotic settings, feature extraction techniques and viewpoints. Our portable active binocular robot vision architecture achieved average recognition rates of 93.5% for fronto-parallel viewpoints and, 83% percentage for anthropomorphic viewpoints, respectively.

1 Introduction

Active robot vision systems are dynamic observers that exploit recovered information from the imaged scene to perform actions and fulfil tasks [7]. Active robot vision systems mainly comprise hard-wired, ad-hoc visual functions that are intended to be capable of robustly exploring a scene and finding objects contained in a database of pre-trained object examples [9] [10]. However, current systems are limited in their visual capabilities and their software modules are crafted according to the robot's specific geometric configuration and hardware components. These limitations constrain the scope of potential applications for such vision systems.

In this paper, we present a portable active binocular robot head architecture that is able to execute *vergence, localisation, recognition and simultaneous identification of multiple target object instances*. In this paper, we focus on the development of a portable architecture while preserving visual behaviours previously reported in [2, 3]. We have chosen the Sensor Fusion Effects(SFX) architecture [16] as the foundation for our portable robot head (Fig. 1). We must point out that *our robot architecture is not an attempt to model the mammalian visual pathway itself*, but it is a functional system that robustly carries out the specific high-level task of *autonomous scene exploration*. To demonstrate the portability of our system, we conducted experiments considering three important variables for any active scene exploration tasks, namely; the hardware used, visual representation and view(s) of the scene. Hence, we present experiments with three different state-of-the-art feature extraction techniques, namely SIFT [12], SURF [8] and KAZE [1] and, different hardware and scene settings.

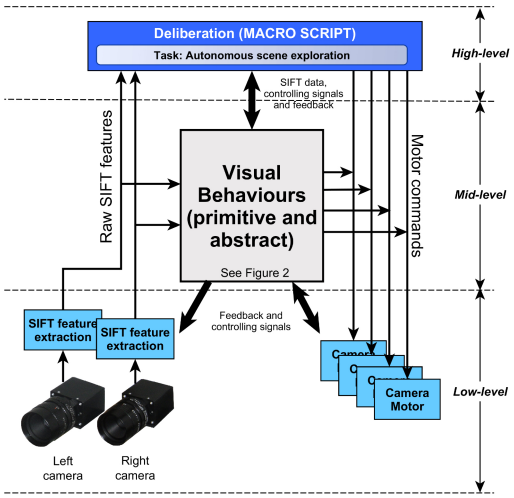

Fig. 1: Our active binocular robot vision architecture.

This paper is organised as follows: Section 2 presents a literature review of current robot vision technologies. Section 3 and 4 presents our robot vision architecture. Finally, Section 5 and Section 6 details the experimental validation of the system and concluding remarks of this paper, respectively.

2 Literature Review

In robotic vision, active vision can potentially offer a sheer amount of information about the robot's environment. Should a visual task becomes ill-posed, the gaze of a robot can be shifted to perceive the scene from a different viewpoint [7]; and therefore a better understanding of the task. Current research in active robot heads has focused on the *"lost and found"* problem [15]. That is, a robot is commanded to search and locate an object in its working environment for exploration tasks [10, 6], manipulation tasks [18, 20] and/or navigation [15].

In an effort to replicate the nature of visual search scan paths [21], researchers have proposed a variety of visual search mechanisms according to the task at hand (e.g. [18, 15, 13]). These heuristic approaches are mainly driven by the outputs of available feature extraction techniques. For example, Rasolzadeh et al. [18] used depth to segment the scene according to the distance between a targeted object and the robot as part of a visual object search heuristic. Likewise, Merger et al. [15] implemented a saliency map that combines intensity, colour and depth features to drive attention, biased by a top-down feature detection based on the MSER feature extractor [14] for object recognition and navigation. Aydemir et al. [6] have recently presented a strong correlation between local 3D structure and object placement in everyday scenes. By exploiting the relationship

between local 3D structure and different object classes, the authors are able to localise and recognise complex 3D objects without implementing specialized visual search routines. Finally, Collet et. al. [10] have proposed an Iterative Clustering Estimation (ICE) algorithm that combines feature clustering along with robust pose estimation. This approach relies on creating sparse 3D models to localise and detect multiple same-class object instances. Advancements in visual search mechanisms have been promising in recent years of which they are not merely restricted to the feature extraction used and rather powered by cognition. For instance, a notable approach proposed in [11] looks at the problem of a robot searching for an object by reasoning about an object and possible interactions with the object. However this robot vision system is limited to one single instance per object class in the scene.

The vision architecture we present, advances the robot vision system described in [2, 3]. That is, we have previously reported an active vision system that is capable of binocular vergence, localisation, recognition [3, 2] and simultaneous identification of multiple target object instances [4]. We structure this initial system as a collection of ad-hoc functions in order to explore autonomously a scene by operating solely with SIFT features. Our system was also constrained to the hardware and, therefore, the limitation of its portability remained an issue. Recent developments in robotic middleware (e.g. the Robot Operating System [17]) technologies have made possible the deployment of hardware independent robotic systems. We thus propose an active binocular robot head architecture that integrates visual behaviours in a parsimonious and generic robot vision architecture based on the Robot Operating System (ROS).

While we do not make explicit use of 3D information in this paper, an explicit goal was to determine if we could reliably maintain binocular vergence of an actuated stereo-pair of cameras while actively exploring a scene. This converged binocular camera configuration supports the recovery of feature locations in 3D and also provide images for stereo-matching for dense 3D range map extraction . This feature underpins visual competences for other robotic applications as demonstrated in [19] where we presented a dual-arm robot manipulating deformable objects using the binocular system reported in this paper.

3 Robot Vision Architecture

As stated before, we have based our active vision system on the hybrid deliberative/reactive *Sensor Fusion Effector* architecture (SFX, [16]). Specifically, the SFX architecture, as implemented, relates how deliberative and reactive modules are interconnected with sensor and actuator functions. Visual behaviours in our architecture implement the configuration of the visual streams in the mid-level of the SFX architecture. This arrangement exploits sensed visual information in order to explore the environment without further reasoning (i.e. the mid-layer *senses and acts* accordingly) while the deliberative layer manages visual behaviours and, consequently, orchestrates the required set of commands to carry out a *high-level* visual task; for instance, manipulation/interaction tasks [19].

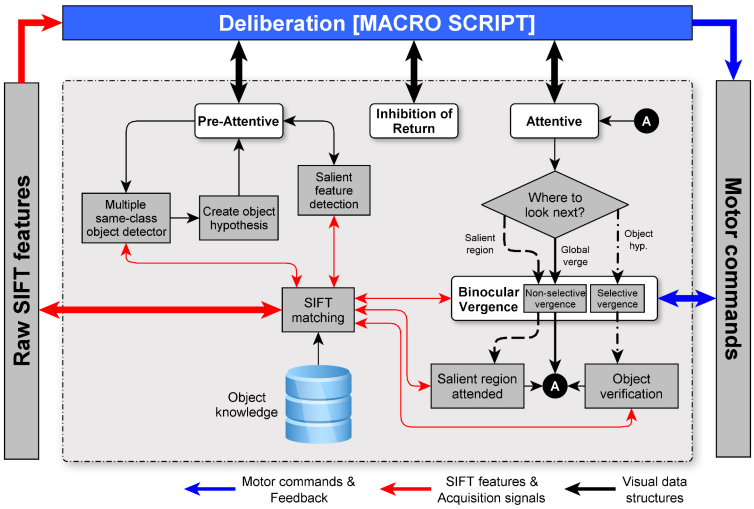

Fig. 2: Internal representation of visual behaviours (Figure 1). White boxes denote abstract behaviours, whereas grey boxes represent primitive behaviours.

Specifically, Figure 1 shows our architecture. The processing levels are classified in terms of their function (i.e. low-level, mid-level and high-level). The corresponding low-level and mid-level functions consist of simple yet effective behaviours that subserve upper-level goals, whilst the high-level functions relate to the intelligence, deliberation and reasoning (out of the scope in this paper).

High-level functions (as observed in Figure 1(a)) specify visual tasks and goals. This layer, this paper, is cast as scripted meta-behaviours (Section 4) that orchestrate the sequential activation of visual behaviours in order to fulfil the task of autonomous visual object exploration.

Low-level and mid-level (Figures 1(a) and 2(b)) integrate a number of *primitive* and *abstract behaviours*. On the one hand, primitive behaviours comprise monolithic methods that only serve a single purpose; i.e. they are simple stimulus-response mappings that transform a collection of sensed information into data structures. On the other hand, abstract behaviours comprise a collection of primitive or other abstract behaviours. Figure 2(b) illustrates the **mid-level** processing architecture that comprises *pre-attentive*, *attentive*, *inhibition of return* and *binocular vergence* visual behaviours previously reported in [2, 3]. Sensor and motor behaviours are decoupled from the mid- and high-level layers. This configuration allows us to maintain visual behaviours that are not constrained to the chosen feature extraction technique and hardware components.

To achieve generic and preserve a modular arrangement within the architecture, we devised an egocentric coordinate system which are not related to the real-world units of the observed environment. The egocentric coordinate map is

Table 1: Pseudo-code of macro script in Figure 1 and 2.

```
Inputs: None
Outputs: List of objects recognised and attended to.
```

```
1:  Generate database
2:  Verge cameras and extract features from the image pair
    (binocular arrangement)
3:  Obtain pre-attentive object and salient hypotheses
4:  Set the saccade number to 1
5:  Loop until possible object or salient hypotheses are not empty
    or no. of saccades is less than a user-defined number
6:     Select an object from the possible obj. hypotheses that has
       the maximum recognition score (see [3])
7:     Verge and attend to the selected object and return features
       from both cameras after verging and the lists of the
       remaining object and salient hypotheses
8:     Update pre-attentive object and salient hypotheses
9:     Inhibit (inhibition of return) new pre-attentively found
       objects w.r.t previous possible object and salient hyps
10:    Saccade no. increments 1
11:  Report objects stored
```

defined as a relative pixel-based map where the frame of reference is established with respect to a *"home"* position of the robot head.

4 Visual Search Task Definition

The high-level layer is defined as a macro-script that specifies the visual search task, controls and schedules behavioural resources in lower layers (ref. [3]), and monitors the progress of the task. In this paper, we define a *pre-attentive-inhibition of return-attentive cycle* in order to allow our system to perform autonomous scene exploration (Table 1). That is, the robot acts according to the sensed visual information and reports recognised object classes stored in database.

By replacing the macro script with a cognitive/intelligent layer, the sequence of behaviours required to convey a visual task can be generated deliberatively thereby removing the fixed-task limitation of the current control scheme. Accordingly, the architecture we describe here has been designed such that a deliberative/cognitive module might replace the fixed script in future modifications of the robot system without altering the underlying visual behaviours.

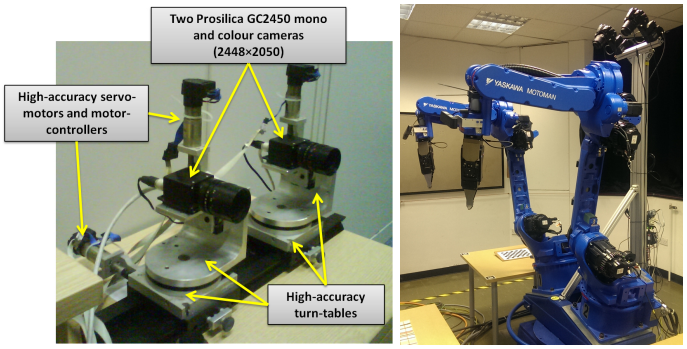

Fig. 3: Left: The Prosiclica robot head exploring the scene. Right: An image of the dual-arm robot featuring the Nikon robot head on top. Additionally, this robot features grippers specifically designed for manipulating clothing [19].

5 Experiments

5.1 Robot Head Hardware and Software Interface

These experiments are designed to validate the portability of our active robot vision architecture in two different scene settings and hardware components. The first active binocular robot head (Figure 3) comprise two *Prosilica* cameras (*GC2450C* and *GC2450*; colour and mono, respectively) at 5 Mega pixels of resolution fitted with *Gigabit Ethernet* interfaces and 4 high-accuracy stepper-motors and motor-controllers (Physik Instrumente). The robot vision architecture is arranged as follows for the latter robot head. Low-level components, namely, image acquisition and motor control modules (Figure 1); are interfaced to a Pentium 4 computer with 2 GB in RAM running under Windows XP and MATLAB R2008a. Whilst, image feature extraction, mid-level and high-level components (Figure 1) are interfaced to a 4-core Intel Xeon (model E5502) with a CPU clock speed of 2 GHz, with 24 GB in RAM running under Windows 7 and MATLAB R2009b. Both computers are interconnected through the local network by means of a collection of network socket functions for MATLAB[1].

The second active binocular robot head (Figure) consists of two Nikon DSLR cameras (D5100) at 16 Mega pixels of resolution. Cameras are mounted on two pan and tilt units (PTU-D46) with their corresponding controllers. This robot head is mounted on a dual-arm robot with anthropomorphic features. Low-level functions where implemented as ROS nodes and interfaced with Matlab 2014a with *pymatlab*[2]. The hardware is interfaced to an Intel Core i7-3930K computer at 3.20 GHz with 32GB of RAM running Ubuntu 12.04 and ROS.

[1] `http://code.google.com/p/msocket/` (verified on 4 March, 2016)
[2] `https://pypi.python.org/pypi/pymatlab` (verified on 4 March, 2016)

Fig. 4: Left: View from the Prosilica robot head's left camera exploring a scene. Right: View of the Nikon-based robot head as viewed from the left camera.

5.2 Methodology

In order to test the robustness and repeatability of our architecture, for both binocular robot heads, we performed 3 visual exploration tasks for each scene, each visual task with a random initial home position. It must be noted that we terminate the visual search task if the robot's pre-attentive behaviour does not find an object within 5 consecutive saccades; i.e. the system is only targeting salient features. This halting criterion has been implemented in order to reduce the execution time while conducting these experiments.

There are three possible outcomes while actively exploring a scene:

- *True positives* comprise all correctly detected and identified object hypotheses where the system is able to centre the hypothesised object in the field of view.
- *False positives* include when the system localises an object hypothesis, but without being able to centre the object in the field of view of both cameras during the attentive cycle or, similarly, an attended object hypothesis does not correspond to the object class in the scene.
- *Not found* comprise the system's failures when an object instance is not detected in the visual search task.

For each robot head, we have arranged scenes comprising a mix of several multiple same-class and different-class object instances, arranged in different poses. We define scene complexity according to the number of similar unknown objects in the scene (i.e. a typical source of potential outliers) and by the degree of background clutter present. We detail below the experimental methodology.

Prosilica Robot Head. We arranged 7 different scenes[3] of differing complexity, based on combinations of 20 known object instances, of 10 different object classes. Figure 4 shows an example of a scene. Objects were placed in arbitrary poses and locations. We have also created a database of the 10 known objects by capturing stereo-pair images of an object at angular intervals of 45° and 60°. These captured images are then manually segmented in order to contain only

[3] All 7 scenes can be accessed at `http://www.gerardoaragon.com/taros2016.html`

(a) Scene 1 (b) Scene 2 (c) Scene 3

Fig. 5: Scenes used for the Nikon robot head. a) Scene 1 depicts less complexity. b) Scene 2, medium complexity. c) Scene 3, most complex scene of the last two.

the object of interest. We have considered two databases in order to measure the recognition performance of our system with different visual knowledge.

Nikon Binocular Robot Head. Scenes for these experiments consist of objects placed on top of a table. The goal is to investigate the response of our active vision architecture to different viewpoints, different feature extraction techniques and hardware components for the sake of portability. With this robot head, we are also able to investigate the effects of having an anthropomorphic robot configuration as opposed to a fronto-parallel configuration as above. Figure 5 shows examples of the scenes we created. Object databases used in these experiments include stereo-pair images of object instances sampled randomly in order to cover the objects' view-sphere by placing the object in isolation on the working table. Each object instance stored in the database is manually segmented.

We therefore arranged 3 different scenes[4] of variable complexity. Each scene is a composition of 14 known object instances observing arbitrary poses and locations, of 9 different object classes. Scene 1 is considered to be the simplest while scene 3, the most complex (Figure 5). We must note that Scene 2 and Scene 3 include flat objects and objects with 3D structure while Scene 1 only comprise objects having 3D structure. In order to effectively understand the response of the system to different feature extraction techniques, each of the three scenes were explored by our system with SIFT, KAZE and SURF features.

5.3 Analysis and Discussion

Investigating all experiments and three randomly starting position for each scene, we can deduce that our active robot vision architecture presents stochastic behaviours. Accordingly, neither robot vision head follows a pre-defined visual scan path but it adapts according to the contents of the scene while exploring the scene. Summary of the outcomes for each robot head are presented as follows.

[4] All 3 scenes can be accessed at `http://www.gerardoaragon.com/taros2016.html`

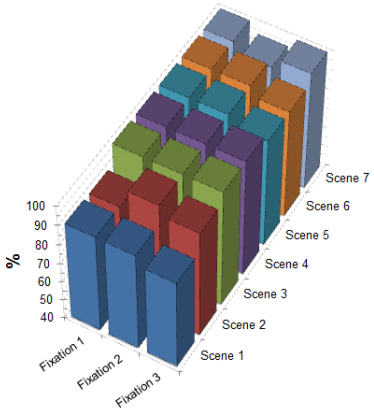

Fig. 6: Overall recognition rate for the visual tasks for the Prosilica robot head.

Table 2: Outcomes for the Prosilica robot head.

Scene No.	Performance (%)	True Positives	False Positives	Not Found	Recover from Failures
1	82	56	5	7	0
2	93	57	1	3	0
3	97	60	2	0	2
4	97	60	2	0	2
5	91	59	5	1	4
6	98	59	0	1	1
7	97	59	1	1	0
Total	**93.5**	**410**	**16**	**13**	**9**

Prosilica Robot Head. Table 2 illustrates the system's recognition rates for all experiments. False positives emerged due to the object feature descriptors matching with unknown objects and, in consequence, these matches were not consistent with the reference object centre in the database while generating object hypotheses pre-attentively (as previously reported in [2]). However, the system recovered from false positives. These results further support the active vision paradigm, since the robot vision architecture is able to recover from these failures while investigating the scene from different views. Thus, the robot is able to locate almost all of the object instances, despite not noticing every object instance present during each pre-attentive cycle.

Nikon Robot Head. From table 3, we can observe that the recognition performance is linked to the feature extraction techniques used. Average recognition rates for SURF, SIFT and KAZE are 60%, 77% and 83% percentage, respectively. SIFT and KAZE, in these experiments, achieved better recognition rates than SURF due to the inherent properties of being "almost" invariant to perspective transformations. It is also worth noting that both SIFT and KAZE

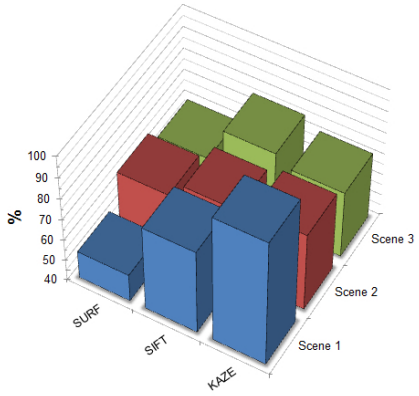

Fig. 7: Outcomes for experiments with the Nikon robot head.

Table 3: Outcomes for the Nikon robot head.

Descriptor	Scene No.	Performance (%)	True Positives	False Positives	Not Found	Recover from Failures
SURF	1	53	16	5	14	54
	2	66.6	26	4	13	65
	3	59.5	25	2	17	64
Total		**60**	**67**	**11**	**44**	**184**
SIFT	1	80	24	0	6	30
	2	74	29	0	10	54
	3	76	32	0	10	59
Total		**77**	**85**	**0**	**26**	**143**
KAZE	1	100	30	0	0	30
	2	76.9	30	0	9	39
	3	71.4	30	0	12	29
Total		**83**	**90**	**0**	**21**	**98**

techniques are less prone to false positives as opposed to SURF. As we described above, our portable active vision architecture was tested using an anthropomorphic configuration where objects are not in similar 2D planes as it is the case from the Proscilica robot head experiments. By comparing Table 3 with Table 2, we can observe a decrease in the performance. That is, 3D structures from an anthropomorphic configuration are more difficult to recognise and, therefore, the robustness of feature descriptions decrease. We can also observe more *recoveries from failures* (last column in Table 3) in these set of experiments. We deduce that this particular configuration introduces challenging geometric transformations that state-of-the-art feature descriptions are still not able to cope with. Hence, the chosen feature extraction has a key role in the overall recognition performance. Nevertheless, our active robot head is able to explore a

scene regardless of hardware configuration, different view point while maintaining acceptable recognition rates.

6 Conclusions and Future Work

We have presented a portable active binocular robot head that integrates visual behaviours in a unified and parsimonious architecture that is capable of autonomous scene exploration. That is, our robot architecture can identify and localise multiple same-class and different-class object instances while maintaining vergence and directing the system's gaze towards scene regions and objects.

Our portable robot vision architecture has been validated over challenging scenes and realistic scenarios in order to investigate and study the performance of the visual behaviours as an integrated architecture. By carrying out a qualitative comparison with current robot vision systems whose performance has been reported in the literature, we argue that our architecture clearly advances the reported state-of-the-art [5, 18, 15, 3, 13] in terms of our system's innate visual capabilities and portability to different environment settings, e.g. multiple same-class object identification and tolerated degree of visual scene complexity. Our architecture is therefore portable enough in order to be adapted to different hardware configuration, feature description and view-points.

In biological systems, it is found that a region in the scene that is sufficiently salient can capture the attention of an observer more than once during a visual task [22, 21]. Our current inhibition of return behaviour, however, has been formulated explicitly to prevent the robot from visiting a previously attended location. We propose to revise this behaviour by incorporating an exponential decay criterion that dictates the mean-lifetime of inhibition of an attended location. The robot would then be able to re-visit a previously attended location, perhaps in the context of a spatial awareness model with a cognitive module.

Acknowledgements This work was partially supported by the Programme Alβan, the European Union Programme (grant number E07D400872MX) and CONACYT-Mexico (grant number 207703).

References

1. Alcantarilla, P.F., Bartoli, A., Davison, A.J.: KAZE features. In: Eur. Conf. on Computer Vision (ECCV) (2012)
2. Aragon-Camarasa, G., Siebert, J.P.: A hierarchy of visual behaviours in an active binocular robot. In: Kyriacou, T., Nehmzow, U., Melhuish, C., Witkowski, M. (eds.) Towards Autonomous Robotic Systems, TAROS 2009. pp. 88–95 (2009)
3. Aragon-Camarasa, G., Fattah, H., Siebert, J.P.: Towards a unified visual framework in a binocular active robot vision system. Robotics and Autonomous Systems 58(3), 276–286 (Mar 2010)
4. Aragon-Camarasa, G., Siebert, J.P.: Unsupervised clustering in Hough space for recognition of multiple instances of the same object in a cluttered scene. Pattern Recognition Letters 31(11), 1274–1284 (Aug 2010)

5. Arbib, M., Metta, G., der Smagt, P.: Neurorobotics: From vision to action. In: Siciliano, B., Khatib, O. (eds.) Springer Handbook of Robotics, pp. 1453–1480. Springer Berlin Heidelberg (2008)
6. Aydemir, A., Jensfelt, P.: Exploiting and modeling local 3d structure for predicting object locations. In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. pp. 3885–3892 (Oct 2012)
7. Ballard, D.H.: Animate vision. Artificial Intelligence 48(1), 57–86 (1991)
8. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. 110(3), 346–359 (Jun 2008), http://dx.doi.org/10.1016/j.cviu.2007.09.014
9. Chen, S., Li, Y., Kwok, N.M.: Active vision in robotic systems: A survey of recent developments. International Journal of Robotics Research 30(11) (2011)
10. Collet, A., Martinez, M., Srinivasa, S.S.: The moped framework: Object recognition and pose estimation for manipulation. The International Journal of Robotics Research (2011)
11. Dogar, M., Koval, M., Tallavajhula, A., Srinivasa, S.: Object search by manipulation. Autonomous Robots 36(1-2), 153–167 (2014)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
13. Ma, J., Chung, T.H., Burdick, J.: A probabilistic framework for object search with 6-dof pose estimation. The International Journal of Robotics Research 30(10), 1209–1228 (2011)
14. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: In British Machine Vision Conference. vol. 1, pp. 384–393 (2002)
15. Meger, D., Gupta, A., Little, J.J.: Viewpoint detection models for sequential embodied object category recognition. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 5055–5061 (2010)
16. Murphy, R., Mali, A.: Lessons learned in integrating sensing into autonomous mobile robot architectures. Journal of Experimental & Theoretical Artificial Intelligence 9(2), 191–209 (Apr 1997)
17. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: Ros: an open-source robot operating system. In: ICRA workshop on open source software. vol. 3, p. 5 (2009)
18. Rasolzadeh, B., Bjorkman, M., Huebner, K., Kragic, D.: An active vision system for detecting, fixating and manipulating objects in the real world. The International Journal of Robotics Research 29(2-3), 133–154 (2010)
19. Sun, L., Aragon-Camarasa, G., Rogers, S., Siebert, J.P.: Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening. In: Robotics and Automation (ICRA), 2015 IEEE International Conference on. pp. 185–192 (May 2015)
20. Sun, L., Aragon-Camarasa, G., Rogers, S., Siebert, J.: Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening. In: Robotics and Automation (ICRA), 2015 IEEE International Conference on. pp. 185–192 (May 2015)
21. Wolfe, B.A., Whitney, D.: Saccadic remapping of object-selective information. Attention, perception & psychophysics 77(7), 2260–9 (oct 2015)
22. Wurtz, R.H., Joiner, W.M., Berman, R.A.: Neuronal mechanisms for visual stability: progress and problems. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 366(1564), 492–503 (2011)