2017-10-27

# Qualitative Review of Object Recognition Techniques for Tabletop Manipulation

Wallbridge, CD

http://hdl.handle.net/10026.1/13087

10.1145/3125739.3132593
Proceedings of the 5th International Conference on Human Agent Interaction
ACM

# Qualitative Review of Object Recognition Techniques for Tabletop Manipulation

**Christopher D. Wallbridge, Séverin Lemaignan and Tony Belpaeme**
Plymouth University
Plymouth, UK
{christopher.wallbridge, severin.lemaignan, tony.belpaeme}@plymouth.ac.uk

## ABSTRACT

This paper provides a qualitative review of different object recognition techniques relevant for near-proximity Human-Robot Interaction. These techniques are divided into three categories: 2D correspondence, 3D correspondence and non-vision based methods. For each technique an implementation is chosen that is representative of the existing technology to provide a broad review to assist in selecting an appropriate method for tabletop object recognition manipulation. For each of these techniques we give their strengths and weaknesses based on defined criteria. We then discuss and provide recommendations for each of them.

## ACM Classification Keywords

I.4.8 Scene Analysis: Object Recognition

## Author Keywords

object detection; pose detection; tabletop manipulation.

## INTRODUCTION

### Context: Near Object Interaction

This paper takes a practical approach to survey the technical landscape on the problem of small object identification and 6D object localisation in a cluttered environment – a context often termed as *object recognition for tabletop manipulation*. Our approach is practical: we consider a typical interaction setup (Fig. 1) where the robot needs to accurately and robustly identify and localise objects in order to manipulate them, communicate about them or reason on their geometric properties and relations. Critically, the object recognition technique needs to be suitable for actual experimental work, including field experiments: it must be reasonably easy to deploy the system in a range of dynamic human environments, without having to rely on expensive or cumbersome physical sensors, or expensive computation. We also take a short to medium horizon: not all techniques we evaluate are commonly available yet, but all have the potential to be robust implementations in the near future.

This paper tries to remedy a lack of information on deployment details in HRI contexts: many traditional assessments do not report on practical considerations. We need to take into account many different factors. For example, how robust is the detection and pose recognition when there are frequent changes to the environment, such as varying backgrounds or changing lighting conditions.

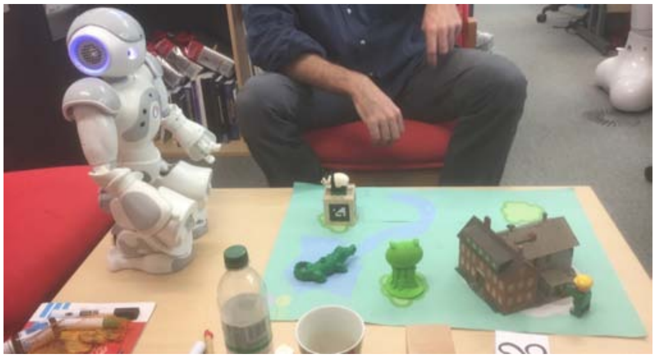

**Figure 1. A close proximity interaction setup, typically found in human-robot interaction and cognitive robotics scenarios. Key scene characteristics are usually constant: relatively small objects (e.g. largest side being less than 10 cm), presence of occlusions, limited working space, and the presence of both textured and texture-less objects.**

In this paper we compare across three families of techniques. The first is techniques that rely on 2D images, from which we track a selection of points. Back projection on these points allow the estimation of an object's 6D position. The second family of methods use 3D templates. 3D objects are compared against a known point cloud to find the position and orientation of an object. The final family relies on techniques that do not use traditional vision techniques, for example RFID technology.

### Surveys on Object Detection

As a cornerstone of many robotic applications, research on object recognition and localisation has been reviewed in numerous past literature surveys. These surveys typically focus on one family of techniques or algorithms, typically using synthetic datasets to quantitatively compare the performances of the state of the art. We summarise hereafter the main findings for each of the localisation techniques.

*HAI '17,* October 17–20, 2017, Bielefeld, Germany

ISBN 978-1-4503-5113-3/17/10. . . $15.00
DOI: https://doi.org/10.1145/3125739.3132593

*Techniques based on 2D correspondences*
When perceptual data consists of camera images, pre-stored templates of objects are often matched against the incoming video stream using 2D correspondence techniques. Li et al. [9] conducted a survey of visual feature detection. In the review they categorise these techniques based on the fundamental principle by which they detect features, such as edge, blob or corner detection. Feature detection methods vary in performance based on the application context, but among them feature based techniques such as A-KAZE, ORB and SURF are popular in object recognition and tracking contexts [5].

*Techniques based on 3D correspondences*
The increased availability and popularity of 3D cameras has driven the need for 3D object matching techniques. Diez et al. [6] performed a qualitative review of 3D registration techniques, in which a mapping is made between 3D images or a 3D templates and an image. They specifically reviewed a variety of detectors and descriptors for 3D registration. Descriptors and detectors attempt to minimise the number of points required before using such brute force techniques to perform accurate identification. Note that while these are used to select salient points, they nearly always end up using iterative closest point (ICP) algorithms, which find corresponding points between a template and an unknown object. The more points that are used, the more accurate the detection is, but using more points has an exponential impact on computational requirements.

*Non vision-based techniques*
Many other reviews also focus on technologies not relying on visual perception. RFID can be used for coarse localisation, and has been shown to have an accuracy of a few centimetres [13]. The techniques used in their review are meant for localisation within a room, while our focus is on techniques that work on the scale of under a metre, for example localising objects on a tabletop. But reduced distance holds potential for increased accuracy, as objects are nearer to the RFID readers. Mautz [10] conducted a wide survey of a number of indoor positioning techniques for a range of applications. Most of the techniques reviewed are localisation for navigation, and are not practical for use in a tabletop situation. However, among the suitable methods identified for the accuracy we require for tabletop recognition was magnetic technology, which is able to reach millimetre levels of precision. Hostettler et al. [8] look at using Anoto positioning technology to localise a robot. They concluded that using a printed pattern that they are able to position a robot with high accuracy and with robustness to lighting and occlusion conditions, the technology was only restricted by the size and quality of the sheets that could be printed with the pattern.

### Approach and Methodology
We compare a number of existing implementations of a wide range of techniques for object and pose detection. We chose a selection of implementations based on availability, ability to process in real-time and that could be considered representative of that technology. Each of these methods was compared against the following criteria:

1. **Degrees of Freedom**: The degrees of freedom that the method is able to measure (position and/or orientation).
2. **Detection Stability**: How stable was the method of detection. Would an object be lost even if nothing was happening, or were false positives generated.
3. **Rotation Invariance**: Is the method able to track the object when it is rotated.
4. **Distance Invariance**: How much does the distance of the object affect the tracking for that method.
5. **Environment Interference**: Is the method able to cope with changes to the background and lighting.
6. **Occlusion**: Can the method detect objects that are being partially occluded by other objects from the perspective of the robot.
7. **Practical Use**: Any additional notes such as extra equipment required that may affect the usability of the system in an experiment.

Each method is briefly described. A table of results provides a side by side comparison of each implementation. Finally we discuss and provide recommendations on each method.

## ASSESSMENT OF OBJECT DETECTION METHODS
Here we briefly describe each method we evaluated and their main weaknesses. Table 1 provides a summary of our results.

### 3D pose estimation from 2D images
These techniques use a standard 2D cameras. From this, image features are extracted that can be used to identify the object. These features can then be used to provide a 3D position by back projecting the 2D points to 3D reference points, using algorithms like ‘perspective-n-point’(PnP) [7].

*Fiducial markers*
Fiducial markers look similar to 2D barcodes that can be printed out or displayed on a screen for detection. Several libraries provide 6D tracking of such markers, like the *chilitags* library [4].

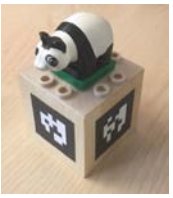

**Figure 2. Object with a fiducial marker, which allows it to be identified and tracked.**

The tags are highly susceptible to occlusion, a small amount is enough to lose tracking. The markers require a flat surface to work, on irregularly shaped objects we get around this by attaching cubes (fig. 2).

*Feature tracking*
Three feature tracking methods were tested using the implementations provided by OpenCV[1]; SURF [2], A-KAZE [1] and ORB [12]. In each case an image is used as a target for the feature detection. These methods are classed as blob detection,

[1] http://opencv.org/

| Method | Degrees of Freedom | Sta. | RInv. | DInv. | Env. | Occ. | Practical Use |
|---|---|---|---|---|---|---|---|
| 2D w/ PnP | | | | | | | |
| Fiducial Markers | 6D | Very High | Very High | High | Very High | Very Low | Markers on flat surfaces |
| A-KAZE | 6D | Moderate | Very High | Low | Low | Moderate | |
| ORB | 6D | Moderate | Very High | Low | Low | Moderate | |
| SURF | 6D | Moderate | Moderate | Moderate | Low | Moderate | |
| Template Matching | 6D | Very High | High | High | Low | Moderate | |
| Deep Learning (Faster R-CNN) | Planar | High | Very High | Very High | Very High | High | High Training Requirement |
| Depth Mapping | | | | | | | |
| ORK | 6D | Very Low | High | High | High | Moderate | RGB-D Camera |
| Realsense SDK | 6D | High | High | High | High | Moderate | RGB-D Camera |
| Non-Vision Based | | | | | | | |
| GaussSense | Planar w/ Rotation | Low | Very High | Very High | Very High | Very High | Sensor Board |
| ePawn | Planar w/ Rotation | Very High | Very High | Very High | Very High | Very High | Sensor Board |

**Table 1. Table showing a summary of the different object detection methods and their performance based on the criteria defined in section 1.3. Sta.: Detection Stability. RInv.: Rotation Invariance. DInv.: Distance Invariance. Env. Environment Interference. Occ.: Occlusion**

which look for areas of pixels that are similar to each other but contrast their surroundings.

All three of these methods struggle with changing backgrounds, and did not handle varying distances well. Besides, computing the backprojection to obtain a 6D pose is generally difficult: as feature trackers select by themselves which features they choose to match, they are not known in advance. This makes it difficult to apply a PnP transformation to recompute 6D coordinates.

### *Template matching*

Template matching, while a relatively old technique, was also considered; we tested using the implementation from OpenCV. An image is used as the target for template matching. This target image is then compared pixel by pixel against an image, and the strongest match is returned as a bounding box.

Multiple target images will be required per object to provide proper 6D pose estimation. Its greatest weakness is to varying backgrounds.

### *Deep Learning*

Deep learning relies on the training of a neural network on a dataset of pictures. Here we used Faster R-CNN [11] to test Deep Learning. We used a pre-trained network[2] that was trained on the PASCAL VOC 2007 image dataset.

The network was unable to detect iconic versions of objects it had been trained on (fig. 3), so training would be required on the specific objects to be used as part of the experimental setup.

This method only provides bounding boxes of the objects, but these cannot be compared against a known object (an object could be small but near the camera or large but far away and we would be unable to determine the exact dimensions). This makes it difficult to provide a 6D estimation.

## 3D pose estimation from 3D sensor data

In recent years RGB-D cameras, which return 3D scene data in addition to a 2D image, have been widely used in HRI. The Microsoft Kinect technology or the Intel Realsense technology have proven particularly popular. Here we evaluate their software in the context of object localisation and pose reading. The techniques that we look do not require more than a tablet or laptop to process the data.

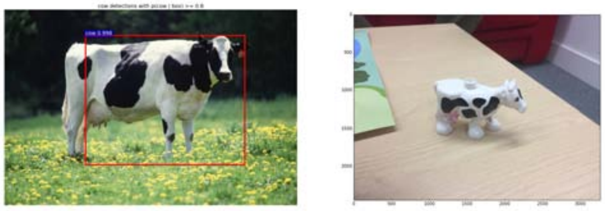

**Figure 3. Images showing two pictures of cows, on the left a real cow that is detected by Faster R-CNN trained on the PASCAL VOC 2007 dataset, on the right an iconic toy cow that is missed.**

### *Planar segmentation and iterative fitting*

We evaluated "Tabletop" from the Object Recognition Kitchen (ORK)[3] implemented using ROS. Tabletop uses planar segmentation to separate the surface of a table and segment objects that are on top. These objects are then compared to a database containing meshes of known objects using simple iterative fitting (related to ICP[3]). This method performed well with different object rotations and scales, and was unaffected by a change in background. However this method generated too many false positives to be considered a stable option for close proximity human-robot interaction scenarios.

### *Intel Realsense tracking*

In the Intel Realsense SDK[4], Object Tracking (C++) for the SR300 was used. This method relies on having a 3D mesh of the object, which it then used for matching. During our investigation we were unable to determine the exact method used by the Intel SDK as it has not been published (see discussion section). Objects were sometimes lost for no apparent reason and would need to be moved for them to be recognised again. This technique is able to handle a small amount of occlusion.

[2] https://github.com/smallcorgi/Faster-RCNN_TF

[3] http://wg-perception.github.io/object_recognition_core/index.html

[4] http://www.intel.co.uk/content/www/uk/en/architecture-and-technology/realsense-overview.html

### Non-Vision Based Techniques

This section details methods that do not rely on the use of cameras, but instead the use of additional equipment.

#### *Magnetic Field sensors*

Magnetic Field sensors use one or more Hall effect sensors to read the position and orientation of a magnetic tag. We evaluated the GaussSense[5] solution, a small and affordable magnet sensor with a high degree of sensitivity. It is able to measure orientation and measures up to 3-4cm away from the sensor. It does however only cover a very small area. Many sensors would be required to cover a larger, the price may then become a consideration, with a 16x16cm board costing $350. GaussSense also requires the use of an Arduino to process the data received. However to distinguish between different tags requires an NFC tag.

#### *NFC solutions*

Several NFC sensors a can be combined into an NFC array, allowing for detection over a larger area. We evaluated the ePawn[6] mat, an NFC sensor board covering a 32x32cm area. The ePawn mat, using a 2D matrix of sensors, can locate a tag with millimetre accuracy. Using two tags in an object allows the calculation of orientation in the plane of an object. Tags themselves are 2cm in diameter so would be able to fit on or inside small objects. Tags only really work well while in contact with the mat. The prototype we evaluated currently costs €1400.

## DISCUSSION AND RECOMMENDATIONS

Of all the 2D vision based techniques fiducial markers were probably the most reliable. However its sensitivity to occlusion means it is unsuitable for a study where the objects are frequently moved around by hand and placed behind other objects. Another challenge is often the attachment of fiducial markers onto objects: curved or irregular objects often prove challenging to attach the markers to. However, fiducial markers might bring benefits not offered by other technologies: the ease of displaying fiducial markers on a screen, or printing out markers, and the high accuracy it can provide, means that it is suitable for calibrating multiple cameras quickly in an experimental setup.

The feature tracking methods (A-KAZE, ORB and SURF) all have issues with dynamic backgrounds, which is an issue when the camera is not static or when subjects in the interaction are in view. It should be noted that the objects being used for this assessment were all relatively simple toys, which lacked rich texture. These methods may perform better on other, more textured, objects, but it may still require combining these methods with other algorithms to get a truly robust detection system.

Template matching, while relatively old, was among the most robust of the 2D methods. To provide a 6D pose estimation however this method will require a lot of templates to compare against. Therefore this method will not scale well with multiple objects. It may be better to use this method to increase the stability of other techniques where it could be used for foreground selection.

The Faster-RCNN that we tested can only provide a bounding box for our objects, this means we cannot get a full 6D pose estimation with this technique alone. However its reliability means that it could be very useful as a foreground selection technique to be used in a pipeline with other methods. Recent research looks into using a CNN that is able to handle 3D pose estimation [14], but it is unlikely that a training set for specific experimental requirements exist as these networks are only just emerging. The process of generating the required training data and then training the network is a process that potentially requires months of work before being usable in an experiment.

The implementation of tabletop in ORK provided too many false positives to be feasible for use in our future studies. However we only tried one camera, the Intel SR300. Other hardware or updates to software drivers may increase performance. By making use of the planar segmentation part of the process it would be possible to subtract the background for use in other detection methods, causing this to no longer be an issue for those methods which struggle with varying backgrounds.

The Intel Realsense SDK performed better with a lot higher stability compared to ORK. However the issue where it would sometimes lose an object while not common is still enough to cause issues in a study. This however is probably the best method available if it is a requirement to track objects while they are being moved. We were unable to find the exact technique that Intel Realsense used, as it has not been published, but due to its performance it was still included in this review. It appears to identify contours in the object before we assume using ICP to match these points to the points of objects stored in the database.

None of the vision based techniques were fully capable of performing the required level of object recognition in a practical tabletop setting. However a pipeline of techniques has the potential to overcome the weaknesses that are shown with just a single method. For instance the 2D techniques could be used to provide a bounding box and classification of the object, allowing a 3D technique to provide precision depth and pose information.

The GaussSense magnetic sensor performs well when tracking a single object. However an NFC module is required to be able to distinguish between multiple objects. For this reason it would be recommended to just use an NFC sensor when using multiple objects.

The ePawn NFC mat is probably the best method reviewed here for use in object recognition with tabletop manipulation. Its downside is that it cannot provide full 6D pose estimation, and the need for additional sensor equipment in the form of a RFID matrix. It is however suitable for many cases where objects need to be tracked, and potential interactions can be shaped around this limitation. NFC also has an advantage of being a known and reliable technique, as it used widely in contactless technology, such as debit cards and key fobs.

[5] http://gausstoys.com/

[6] http://epawn.fr/

## ACKNOWLEDGEMENTS

This work was supported by the EU Horizon 2020 L2TOR project (grant 688014) the and EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

## REFERENCES

1. Pablo F Alcantarilla and T Solutions. 2011. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell* 34, 7 (2011), 1281–1298.
2. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. *Computer vision–ECCV 2006* (2006), 404–417.
3. Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Robotics-DL tentative*. International Society for Optics and Photonics, 586–606.
4. Quentin Bonnard, Séverin Lemaignan, Guillaume Zufferey, Andrea Mazzei, Sébastien Cuendet, Nan Li, Ayberk Özgür, and Pierre Dillenbourg. 2013. Chilitags 2: Robust Fiducial Markers for Augmented Reality and Robotics. (2013). `http://chili.epfl.ch/software`
5. Hsiang-Jen Chien, Chen-Chi Chuang, Chia-Yen Chen, and Reinhard Klette. 2016. When to use what feature? SIFT, SURF, ORB, or A-KAZE features for monocular visual odometry. In *Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on*. IEEE, 1–6.
6. Yago Diez, Ferran Roure, Xavier Lladó, and Joaquim Salvi. 2015. A qualitative review on 3D coarse registration methods. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 45.
7. Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.
8. Lukas Hostettler, Ayberk Özgür, Séverin Lemaignan, Pierre Dillenbourg, and Francesco Mondada. 2016. Real-time high-accuracy 2D localization with structured patterns. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 4536–4543.
9. Yali Li, Shengjin Wang, Qi Tian, and Xiaoqing Ding. 2015. A survey of recent advances in visual feature detection. *Neurocomputing* 149 (2015), 736–751.
10. Rainer Mautz. 2012. Indoor positioning technologies. (2012).
11. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
12. Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2564–2571.
13. T Sanpechuda and L Kovavisaruch. 2008. A review of RFID localization: Applications and techniques. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*, Vol. 2. IEEE, 769–772.
14. Paul Wohlhart and Vincent Lepetit. 2015. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3109–3118.