Proceedings of the Post-Graduate Conference on Robotics and Development of Cognition
10-12 September 2012, Lausanne, Switzerland

Towards Modular Spatio-temporal Perception for Task-adapting Robots

Zoltan-Csaba Marton, Florian Seidel, Michael Beetz

Intelligent Autonomous Systems, Technische Universität München, Munich, Germany
{marton, seidelf, beetz}@cs.tum.edu

Abstract

In perception systems for object recognition, the advantage of multiple modalities, of combining approaches, and several views is emphasized, as they improve accuracy. However, there are great variances in the implementation, suggesting that there is no consensus yet on how to approach this problem. Nonetheless, we can identify some common features of the methods and propose a flexible system where existing and future approaches can be tested, compared and combined. We present a modular system in which perception routines can be easily added, and define the logic of making them work together based on the lessons learned from different experiments.
Index Terms: robotics, multi-cue vision, machine learning

1. Motivation

Autonomous agents working in human environments have a huge variety of objects to deal with, and some of them present special problems (texture-less, transparent, etc). There are multiple approaches that have been shown to be able to segment, detect, categorize and/or classify some of the objects such robots might encounter. There are, however, inherent limitations in these approaches, and there is no robust and large-scale solutions yet [1]. As each perception method captures only some aspect of the objects, the situation is similar to the old story about the six blind men trying to describe an elephant based on a single touch. Clearly, a correct combination of different sensor modalities, segmentations, features, classifiers would improve results. Additionally, in [2] it is argued, that a cognitive agent needs to be embodied to gather experiences, and presents different paradigms on how to approach the learning and grounding of new information. Similar ideas are discussed in [3] as well, where the task and environment adaptation of a robot improves its capability to perceive objects.

In this work, we focus on taking advantage of exploration capabilities of the robot, and the fact that a high-level task specification is typically available. Therefore we propose a system that can take advantage of the fact that only some objects are probable to be at different places in the close surrounding of the robot, and of these ones, only some are relevant for the task at hand. Different perception methods can then be activated (or tuned) and combined, in order to improve detection rates. Additionally, multiple observations over time can be incorporated to obtain higher quality results. In short, the main propositions of this paper for a perception system are as follows:

- common input-output defined for segmentation and detection methods,
- support for consecutive or parallel methods to correct or support each-other in a probabilistic framework,
- enable the specialization of each method to a subset of objects and to group objects into categories,
- incorporating information from multiple views to disambiguate complex cases.

To support our approach, we evaluate these principles, and:

- show the advantage of combining different cues,
- evaluate different ensemble methods and discuss their benefits and drawbacks,
- describe our practical solutions to increase the robustness and accuracy of perception systems,
- present proof-of-concept experiments.

After an overview of the related work, we will outline the basis for our proposal in Section 3, followed by the details of a multi-cue perception system in Section 4. As it will be detailed, the modular combination of task-adapting perception routines performing spatio-temporal integration of multiple modalities holds great potential for the development of robust computer vision. We argue that a deep integration of various levels of a cognitive architecture will be required, and present the connections we found to be most important in our experiments.

2. Related Work

Inspired by earlier work based on developmental psychology, object categorization using multiple modalities is explored in [4] and the advantage of accumulating information over time is shown. While sychological findings do suggest that a single sensory modality is often not enough, they leave out the most descriptive modality, vision, and focus on proprioceptive and auditory feedback [5].

In [6] the authors validate the use of different visual modalities, showing that color-based cues are more important for instance recognition, while geometric ones are better suited for categorization, and that their combination improves on both.

Existing perception systems that use multiple modalities for object detection, either combine these in a fixed feature [7] or use them in a fixed framework [8]. Selecting only relevant features for a specific task was explored in [9], but in a sequential framework with a fixed order of features/modalities. Here we propose a parallel architecture with a heuristic decision on which perception primitives should be applied to identify different objects, and with an incremental merging and verification step to provide the final result.

Systems that use validation of the detections through geometric consistency relied on a single modality so far [10, 11], however the advantage of scoring or voting for different solutions is an important lesson that we incorporated in the system.

There is growing evidence that human vision combines top-down (concept driven) and bottom-up (data driven) approaches [12], thus extending classification systems with context information is a natural way of increasing performance. In our framework we use the prior distribution over the possible objects/locations (and the known object models) as the context.

Most of the perception systems rely either on color/black&white camera (e.g. SIFT [13]) or 3D (e.g. VFH [14])information, although image processing techniques can be applied on different image sources as well (e.g. thermal cameras). There are approaches that combine geometry and color descriptors, but properly balancing these two is not straightforward as discussed in [15].

3. Experimental Support

Some aspects of the proposed solution have been verified already in different experiments. The following subsections give details on the evaluation of some of the natural ways how object perception results can be improved.

3.1. Multi-Modal Perception

Combining multiple sensor modalities to improve detection can be done in general either by combining multiple features in a single classification pipeline or by separate processing pipelines for each modality, whose results are combined. The former approach is pursued in [6], where a combination of visual and depth cues is used. We explored the latter approach in [9], highlighting the limitations of the different sensors, and exploiting that not all features need to be check if there is a subset of them that uniquely describes the object. In this work, we present our approach for combining the results of different modules by forming ensembles, as discussed in the following subsection.

3.2. Ensemble Learning

We evaluated the accuracy of standard off-the-shelf classifiers, trained on image-based and 3D features, and ensembles of such classifiers on the large RGB-D object dataset from [16]. As visual features we used SURF [17] and Opponent SURF with a Bag of Features approach and VFH [14] and GRSD-, the geometric part of VOSCH [15], as geometric features. Our interest lies predominantly in simple, non-parametric ensemble methods, since such simple ensembles can endow the proposed system with the required modularity. Hence, the goal was to investigate how simple, non-parametric ensemble methods compare to more sophisticated but parametric classifiers and ensembles.

As a benchmark we considered the the task of identifying the category to which an object belongs for all of the 300 objects and 51 categories in the dataset. All the objects are seen during training time and half of the over 200,000 scans in the dataset are used for training the classifiers. A quarter is used for evaluation and another quarter as hold-out data to estimate the accuracy of the ensemble methods.

We tested SVM and boosted decision trees (AdaBoost) as classifiers, and different voting based methods and stacking for merging their results, as these were suggested in the literature as promising approaches [18, 19]. Classifiers trained on the concatenation of all the features are used as a baseline to which the performance of the ensembles is compared (see Table 1).

Table 1: *Error rates for single features and the concatenation of all features – linear SVM (top) and AdaBoost (bottom).*

VFH	GRSD-	SURF	O.SURF	All
0.133	0.409	0.281	0.301	0.031
0.149	0.435	0.360	0.361	0.0991

After trying several weightings for the voting methods, the best one was found to be the weighting with the estimated class accuracy. For stacking we used real AdaBoost as level-0 classifiers and real AdaBoost, LogitBoost and Gentle Boost as well as linear SVM and SVM with Radial Basis Function kernel as the level-1 classifier, and found Gentle Boost to give best results.

Table 2: *Voting vs tacking for ensembles of single features*

Base classifier	Voting error rate	Level-1 classifier	Stacking error rate
SVM-Linear	0.100	GentleB	0.054

As shown in Table 2, combining different cues is advantageous, and (while more tests could be made) it seems that concatenating the features outperforms the simple weighted voting and the learning based stacking approach. Nonetheless, both approaches improve the result over those of the best single feature, and we found that using pairwise concatenations of features the error rates can be lowered even below that of the classifiers trained on the concatenation of all the features (see Table 3). This suggests that with the right feature combinations and weighting factors, voting could be a great solution as well – increasing the modularity of the perception system.

Table 3: *Stacking with classifiers of single + double features.*

L-SVM	RBF-SVM	AdaB	LogitB	GentleB
0.031	0.065	0.02	0.019	0.019

3.3. Spatio-Temporal Integration

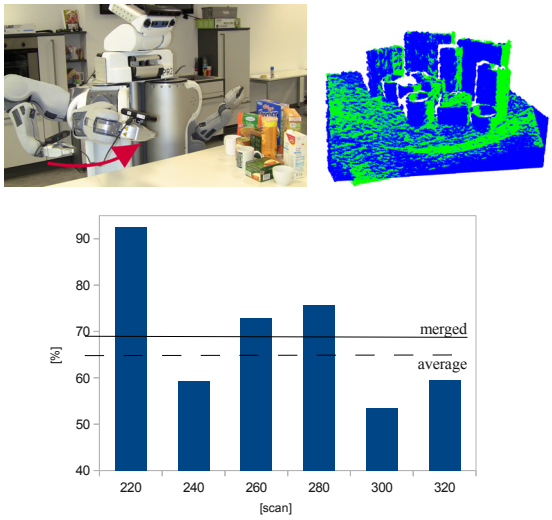

Figure 1: *As the camera is moved (left), multiple frames can be captured that cover different parts of the objects in the scene (right), increasing the overall classification accuracy (bottom).*

We showed the advantage of merging the object detection results from multiple 3D scans in a voting framework previously in [11]. There, we also proposed the use of multiple segmentations of the same input to be merged in the same manner. This approach is employed for image segmentation as well, showing improved results. In Figure 1 the same idea is applied

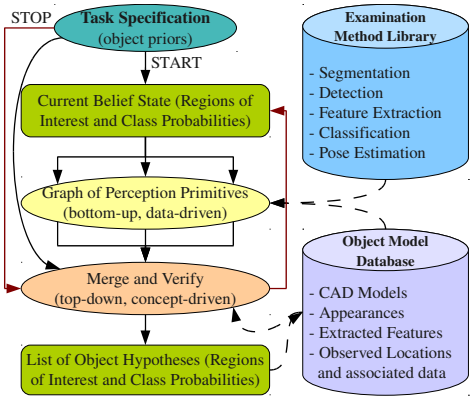

Figure 2: *Basic setup of the proposed system for iterative refinement of object hypotheses by multiple methods according to a task specification.*

for detecting 6 commonly occurring shapes of household objects (a grouping of data from [16] based on [9]), where 3D volumes obtain votes from scans taken from different angles.

3.4. Class Hierarchies

In order to match the perception capabilities of humans, the authors in [20] advocate that searching for predefined templates is not enough, and that recognition of new exemplars of known categories have to be facilitated. On this premise, in [9] we used geometric cues for categorization and visual cues for instance classification. We also reported on the improvement in accuracy of over 10% when the geometric categorization is allowed to work with "internal" categories. This suggests that an unsupervised classification level followed by a mapping to human-defined labels, as in [11], enables the classifiers to tune themselves to the specific feature space used.

4. Proposed Solution

Our proposed solution to integrate the approaches supported in Section 3 for a modular, multi-cue perception system that takes advantage of the robot's exploration capabilities is exemplified in Figure 2 (as a generalized extension of the system presented in [9]). It builds on the lessons learned form previous experiments by the authors and others, and on many discussions from people involved not only in perception, but also high-level planning, manipulation and knowledge engineering for example.

4.1. Regions of Interest and Poses

Most related systems from literature are either doing segmentation or classification (or both at once), but in both cases a region of space is observed, and hypotheses are given about what objects it, or parts of it, contains. A segmentation routine for example breaks large regions up into smaller ones, and assigns to each of them a non-informative prior, i.e. from the point of view of the method each segment can contain anything. Subsequent processing (classification) steps then refine these possibilities. Template matching methods for example do both steps at once, by returning possible (scored) positions in which an object could be in the scene.

Therefore, we propose the use of volumes of space, or *regions of interest* (ROIs) as the basic input *and* output data for object perception methods. These can be for example the hulls

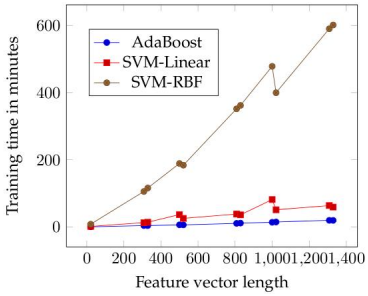

Figure 3: *Feature vector length vs. Training Time (20 classes)*

of clusters for 3D data, or the estimated volumes of image pixels. These, and the associate probabilities of given objects being contained in it, are received and updated by the perception methods, and can be used to merge information coming from different sensors, and different views.

4.2. Task-adapting Perception Primitives

Initially, the system would start off with the complete workspace of the robot as its region of interest, with the different priors for the occurrence of the possible objects assigned to it. The list of these objects and their prior probability can then be considered by the different methods, and when summarizing their results.

We call all the segmentation, detection, fitting and classification methods *perception primitives*, as they are the different modules the system is build of. They can use different sensors, extract various features, apply different recognition methods, and have only to respect the aforementioned input and output in order to be part of the "ensemble". Classification methods such as those based on nearest neighbors, are easy to be re-trained, and this allows simple integration of new data as well. However, with the addition of more and more classes, the accuracy can drop – this can be avoided by taking advantage of the known *task specification* (i.e. list of possible objects and list of sought objects). Similarly, the accuracy of other classifiers deteriorates with the increase in the number of classes (see Tables 4,5 and those in Section 3.2), something that can be alleviated by task and environment specialization.

Table 4: *Error rate for single/concatenated features, 20 classes.*

Classifier	VFH	GRSD-	SURF	O.SURF	All
SVM-linear	0.081	0.270	0.154	0.163	0.0188
SVM-RBF	0.050	0.202	0.098	0.105	0.0172
AdaBoost	0.087	0.293	0.254	0.202	0.0544

Table 5: *Stacking with classifiers trained on single + double features, for 20 classes.*

L-SVM	RBF-SVM	AdaB	LogitB	GentleB
0.013	0.013	0.014	0.014	0.012

Not all classes are as fast to be re-trained as nearest neighbors though, as shown in Figure 3, but methods like locally weighted logistic regression [21] could be used to avoid retraining by adjusting only the weighting of the examples.

4.3. Combining Cues

Since each perception primitive refines the result of its input, the ROIs are trimmed down (if necessary) and the class probabilities accumulated. In the *merging* step all the results can be united through ROI unification, and a decision can be made by an ensemble method. Subsequent sensor readings can be accumulated using the same procedure, and the object hypotheses and their poses can then be verified if they match the data as in [11]. Accumulating or comparing object poses is more complicated, but a scored list of poses can also be maintained, and checked against the accumulated data in the given volume. Another approach to obtain 6DOF pose directly from camera images is to project CAD models of objects to the image and search for good edge responses. However it is unclear how these methods scale to handling very large number of objects.

5. Initial Demos and Discussion

Proof-of-concept demonstrations of the presented approach were made during the 2nd BRICS Research Camp "From 3D sensing to 3D models" (`www.best-of-robotics.org/2nd\_researchcamp/MainPage`) and the public 2011 CoTeSys Fall Workshop (`www.youtube.com/watch?v=DTaeWITW1kI`). Here, a region of interest is provided by the task executive using the known environment model along with the list of possible objects to be detected. The different detection, classification and model fitting methods decide for each request to activate or not based on the objects to be detected and if they have models for those. Different 2D and 3D methods are chained in order to produce the final result, i.e. list of object locations and locations/poses/models. The task executive then interprets the results, decides on the next action to be taken (which could be repeating a failed procedure) and triggers a new task if necessary (e.g. using the bounds of the detected cutter board to detect the slice of bread). As the number of integrated perception primitives increases, and multiple routines for performing the same task become available, the theoretical consideration presented in this paper become more and more important. The presented approach for taking advantage of multiple sources of information by a modular system proved to be useful and scalable in our initial experiments implemented in ROS (`ros.org`). We are confident that the robustness suggested by the supporting experiments will be of great use for integrating our perception system into a cognitive architecture with similar design philosophy, e.g. based on [22].

6. Acknowledgements

This work was supported by the DFG excellence initiative research cluster CoTeSys (www.cotesys.org). The author would like to thank Team2 from BRICS2: Leif Jentoft, Andre Ückermann, Kyle Strabala and Moritz Tenorth; and the IAS team, especially Nico Blodow, David Gossow, Ingo Kresse, Alexis Maldonado, and Dejan Pangercic, as well as Asako Kanezaki and Ferenc Balint-Benczedi.

7. References

[1] D. Kragic and M. Vincze, "Vision for robotics," *Foundations and Trends in Robotics*, vol. 1, no. 1, pp. 1–78, 2009.

[2] D. Vernon, "Cognitive vision: The case for embodied perception," in *Image and Vision Computing*. Elsevier, 2005.

[3] I. Horswill, "Integrating vision and natural language without central models," in *In Proceedings of the AAAI Fall Symposium on Embodied Language and Action*, 1995.

[4] J. Sinapov and A. Stoytchev, "Object category recognition by a humanoid robot using behavior-grounded relational learning," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[5] D. Lynott and L. Connell, "Modality exclusivity norms for 423 object properties," *Behavior Research Methods*, vol. 41, no. 2, pp. 558–564, 2009.

[6] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *Proc. of Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[7] H. Nakayama, T. Harada, and Y. Kuniyoshi, "AI Goggles: Real-time Description and Retrieval in the Real World with Online Learning," *Computer and Robot Vision, Canadian Conference*, vol. 0, pp. 184–191, 2009.

[8] J. F. Georg Arbeiter and A. Verl, "3D Perception and Modeling for Manipulation on Care-O-bot 3," 2010.

[9] Z. C. Marton, D. Pangercic, N. Blodow, and M. Beetz, "Combined 2D-3D Categorization and Classification for Multimodal Perception Systems," *The International Journal of Robotics Research*, 2011.

[10] M. M. Torres, A. C. Romea, and S. Srinivasa, "MOPED: A Scalable and Low Latency Object Recognition and Pose Estimation System," in *Proceedings of ICRA 2010*, May 2010.

[11] O. M. Mozos, Z. C. Marton, and M. Beetz, "Furniture Models Learned from the WWW – Using Web Catalogs to Locate and Categorize Unknown Furniture Pieces in 3D Laser Scans," *Robotics & Automation Magazine*, vol. 18, no. 2, pp. 22–32, 2011.

[12] J. P. Frisby and J. V. Stone, *Seeing: The Computational Approach to Biological Vision*. MIT Press, 2010, ch. 8: Seeing Objects, pp. 178–179.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3d recognition and pose using the viewpoint feature histogram," in *Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 2010.

[15] A. Kanezaki, Z.-C. Marton, D. Pangercic, T. Harada, Y. Kuniyoshi, and M. Beetz, "Voxelized Shape and Color Histograms for RGB-D," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on Active Semantic Perception and Object Search in the Real World*, San Francisco, CA, USA, September, 25–30 2011.

[16] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Proc. of Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[17] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In CGIV*, 2008, pp. 404–417.

[18] L. Lam and C. Y. Suen, "Optimal combinations of pattern classifiers," *Pattern Recognition Letters*, vol. 16, no. 9, pp. 945–954, 1995.

[19] J. Sill, G. Takács, L. Mackey, and D. Lin, "Feature-weighted linear stacking," *CoRR*, vol. abs/0911.0460, 2009.

[20] S. Dickinson, "The evolution of object categorization and the challenge of image abstraction," in *Object Categorization: Computer and Human Vision Perspectives*, S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr, Eds. Cambridge University Press, 2009, pp. pp 1–37.

[21] K. Deng, "Omega: On-line memory-based general purpose system classifier," PhD Thesis, Carnegie Mellon University, Tech. Rep., 1997.

[22] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty, "Building Watson: An Overview of the DeepQA Project," *AI Magazine*, vol. 31, no. 3, pp. 59–79, 2010.